PROFESSIONAL DEVELOPMENT IN LANGUAGE TESTING AND ASSESSMENT: A CASE STUDY OF SUPPORTING CHANGE IN ASSESSMENT PRACTICE IN IN-SERVICE EFL TEACHERS IN THAILAND

Bordin Chinda, M.A.

Thesis submitted to the University of Nottingham
for the degree of Doctor of Philosophy

November 2009

Abstract

This longitudinal qualitative study concerns the investigation of the impact of a professional development (PD) programme conducted at an English department in Thailand. The PD programme was carried out as a series of nine in-service workshops with five non-native English as a Foreign Language (EFL) teachers in the English Department. The workshops aimed to provide these teachers with theoretical and practical understanding of performance-based language assessment with a focus on the rating process. In the investigation of the impact of the PD on these teachers, individual and focus group interviews were used as the research methods. From the analysis of the data, guided by Grounded Theory, the findings show that the PD programme had a positive impact on the teachers who participated in the workshops. These teachers have become aware of their rating styles, established their own consistent rating styles, become confident when rating students' performances, become critical to the assessment practices, realised roles of teachers in assessment, and recognised possibilities of changes in assessment. In other words, they have become more self-consistent when rating their students' performances and they have become more critical to the assessment being used in the department. The insights gained from this research pose the implications for professional development, indigenous rating criteria and collaborative action research.

Acknowledgments

This thesis would have never been completed without constant support and encouragement from my supervisor, Professor Liz Hamp-Lyons. Her wisdom, patience, stimulating suggestions, encouragement and friendship are the continual motivation that energises this work from start to finish.

I am particularly grateful to the Royal Thai government for a full funding during all four years of my PhD study. My thanks also go to the Educational Testing Service (ETS) for facilitating the work through the Small Grants for Doctoral Research in Second or Foreign Language Assessment. I am also grateful to the English Department, Chiang Mai University, for allowing me to take the study leave and for their full cooperation in my data collections.

Above all, this research would not have been possible without the commitment of a large number of individuals: Catbandit, Papone, Songsri, Tanya, and Wanwisa, who dedicated the valuable time to participate in the main study. My sincere thanks also go to Arkom, Ronnie, Pawida, Muun, and Wawan, the teachers who participated in the pilot study. I also owe debts of gratitude to my friends, too many names to mention here, who gave me much support and encouragement in the past four years.

Finally, this thesis is dedicated to my father and mother. The memories of my childhood are powerful sources of courage and determination that have driven me through the difficult times I have had in my life (including this PhD study!). Especially my mother who always believes in me – without her, I would not have come this far.

Table of Contents

List of Tables

List of Figures

1 Introduction

1.1 Background to the Research

This study focuses on the development of Thai teachers who teach English as a Foreign Language (EFL) working in the English department, Chiang Mai University, Thailand, where I was working before embarking on this research project. My motivation for doing this project was conceived in 2002 when the Ministry of University Affairs (now the Commission on Higher Education) announced a reform of English Language Teaching (ELT) and learning in Thai higher institutions, in particular, on the compulsory General English Education curriculum in the response to the revised National Education Act in 1999. In 2002, I was the coordinator for one of the fundamental English courses. My responsibility, with other coordinators of other courses, was to develop new courses according to the goals and standards prescribed by the Ministry of University Affairs. However, as I did not have any background in education or applied linguistics, and as I was a junior staff member, I had to follow the guidelines suggested by the senior staff members.

When the new Foundation English (FE) courses were implemented in 2003, I was appointed the coordinator of FE 1. By the end of the first year of implementation, I realised that there were many problems with the course, especially the assessment (which – for the first time for the FE courses – included performance-based assessment). Issues of assessment have always been a major problem for the department, but there had not been any substantial or effective attempts to solve these problems. At that time, no one in the department had the necessary expertise to be able to solve these problems. Therefore, I decided to carry out this project to understand the causes of these problems in assessment and the solutions to the problems.

The research, which was a longitudinal study, began with a pilot study which aimed to try out research methods and to understand the nature of the problems of assessment in the department. This three month study revealed that the major problems were the diversity of the knowledge and practices of the teachers who participated in the study on the assessment, and especially the rating criteria. In addition, I concluded that a qualitative research approach would be most appropriate for the main study. After reviewing the related literature and many intense supervision meetings, I decided that to solve the problems, in this context, a professional development (PD) programme in language assessment would possibly be the best solution.

Therefore, in the main study (the second phase of the study) I carried out nine PD workshops for teachers in the Department, in which five teachers participated. At the same time, I collected qualitative data on impact of the PD on these teachers. The findings revealed that the PD had a positive impact on these teachers. To validate the findings, I conducted a follow-up study to further investigate this impact. This longitudinal study; pilot study, main study and follow-up study, spanned over a period of a year and a half.

1.2 Outline of Thesis

This thesis is divided into two parts. Part 1, which includes Chapters 1 to 5, provides background information on the whole of the research project, including the theoretical framework and the literature review that underpins the study, as well as the research methodological considerations necessitated by the study, the process of this longitudinal qualitative research, and the course of data analysis. The second part, consisting of Chapters 6 and 7, presents the findings of the study and the discussion on the findings.

The conceptual part of this thesis (Chapters 2 to 4) describes the theoretical foundations of the study. Chapter 2 includes the literature review on performance-based language assessment - covering general concepts of performance-based assessment, assessment criteria and the rating process. Chapter 3 provides the review of PD in general education and then language testing and assessment. This chapter also investigates concepts in teacher change; including studies in teacher change in general education and language testing and assessment (washback/impact study), research in PD in relation to teacher change, and innovation theories.

The aim of Chapter 4 is to provide a brief background into the Thai research context of this study. I offer a brief historical overview of English language education in Thailand, and outline a number of challenges pertaining to language assessment faced in Thailand. I also introduce the FE courses offered by the English Department, Chiang Mai University - the focus of the present study and where the study was conducted. The second part of Chapter 4 reports the findings of the pilot study. The findings consist of two parts: first, the findings from the observations of the department's general practices in assessment, and second, the findings from the case study of five teachers. The findings of the pilot study were used as the justification and implications for the main study.

The research methodology is presented in Chapter 5. The first part of this chapter describes the research methodology of the main study. In this part, I explain how qualitative research design was employed in the study, and, explain in detail the methods used for data collection. The second part explains the research processes of the different phases of the main study. The research process section consists of the results from the investigation of rater behaviours (which was done in preparation for the main study), purposes of the main study, research questions, data collection process, and participant profiles. In addition, I also outline the procedures I adopted in ensuring the quality and ethical issues of the study. The last part of this chapter

describes the data analysis. In this part, I explain how Grounded Theory, along with its fundamental concepts, was used in the interpretation and analysis process of the data. This part also outlines the procedures employed in the analysis of the data.

The second part of the thesis, the actual data analysis, is divided into 2 chapters (Chapters 6 and 7). In Chapter 6, I offer the brief outline of the PD workshop and the overview of the data from the main study and the follow-up study. From the analysis of the interview data with five teachers, who participated in the PD workshop, four main themes emerged: thinking about assessment, thinking about rating criteria, thinking about the PD, and reported assessment practices. Thus, in this part of the thesis, the data of each teacher is categorised into four sections following these themes. The data from the follow-up study is also presented in the same manner, though each theme is not divided into different section.

Chapter 7 presents the discussion of the data presented in Chapter 6. In this chapter, I firstly explore the changes of five individual teachers as the results from participating in the PD. Four of the five teachers exhibited changes in deconstructing, establishing or changing their rating styles; in realising the roles of teachers in assessment and especially in rating processes; by becoming critical of their own and the Deptment's past and present assessment practices and seeing the possibilities for change; by realising the role of rating criteria and teacher-raters in the rating process and becoming more confident in their ownabilities as raters. The second part of the chapter offers the overall discussion of the impact of the PD on these four teachers; including teachers becoming more intra-rater reliable in their ratings, and becoming critical to the assessment. In the third part, I provide a discussion on the participant who was resistant to change.

Finally, in the Conclusion, I outline some implications of this study for a professional development programme, empirically derived indigenous rating criteria and collaborative action research.

2 Performance-based Assessment

In this chapter, I firstly introduce the debates in language testing and assessment paradigms: traditional testing vs. alternative assessment. After that I explore the fundamental concepts of performance-based language assessment followed by the main characteristics of performance-based assessment: assessment criteria and raters. In the assessment criteria section, I include the discussions on rating scales, analytic vs. holistic scales, and approaches in designing rating scales. Issues of rater characteristics and variables, and rater training are elaborated in the final sections.

2.1 Language Testing and Assessment Paradigms: Traditional vs. Alternative

With the arrival of communicative language teaching, language testing and assessment has also shifted to focus more on the performances of students rather than merely discrete point items of traditional testing (for a review of history of language testing and assessment, see Spolsky, 1995, 2008). Traditional testing emphasises 'the rank ordering of students, privileges quantifiable data for isolated, individual test performances, and in general promotes the idea of neutral, scientific measurement as the goal of educational evaluation'; whereas, the 'alternative assessment' is based on 'an investigation of developmental sequences in student learning, a sampling of genuine performances that reveal the underlying thinking processes, and the provision of an opportunity for further learning' (Lynch 2001a, pp. 228 - 229). In addition, Lynch (ibid.) also reports that in traditional testing, the testing and teaching are separated activities conducted by separate groups of people of which the students have no access to the criteria and a single score is usually reported. On the other hand, in the alternative assessment, assessment and teaching are integrated with active participation of the students as part of the process of developing assessment

criteria and standards. In other words, they are two different cultures. Alderson and Banerjee (2001, p. 228) define alternative assessment as:

> assessment procedures which are less formal than traditional testing, which are gathered over a period of time rather than being taken at one point in time, which are usually formative rather than summative in function, are often low-stakes in terms of consequences, and are claimed to have beneficial washback effect. (For the discussion of washback, see Section 3.2.1.)

However, it should be noted that the term 'alternative assessment' has been defined differently by different scholars, and different terms have been used to refer to the same concepts. Other terms include authentic assessment, performance-based assessment, continuous assessment, on-going assessment, to name a few. For the purpose of this thesis, the term 'alternative assessment' and 'performance-based assessment' are used interchangeably.

Furthermore, Lynch (2003, p. 5) identifies different characteristics of alternative assessments:

- assessment practices are considered as integral to teaching;
- students are made active participants in the process of developing assessment procedures, including the criteria and standards by which performances are judged;
- both the process and the product of the assessment are evaluated; and
- the reporting of assessment results is done in the form of the qualitative profile rather than a single score or other quantification.

In the same vein, Brown (1998) suggests 'new ways' of assessing students such as portfolios, journals, logs, conferences, self-assessment, peer assessment, group work, and pair work. He points out that these assessment activities, which are different from tests, are integrated thoroughly into ordinary classroom activities. They, in addition, do not 'stand out as different, formal, threatening, or interruptive'. These ways of assessment, he adds, provides a way of 'observing or scoring the students'

performance and giving feedback in the form of a score or other information … that can enlighten the students and teachers about the effectiveness of the learning and teaching involved' (p. vi).

On the other hand, Brown and Hudson (1998) disagree with the alternative assessments on the basis that this approach regards credibility, auditability, multiple tasks, rater training, clear criteria, and triangulation of any decision-making procedures as ways to improve the reliability and validity of assessment procedures. They assume that using only these methods without the formal reliability and validity would result in 'irresponsible decision making'. Norris, Brown, Hudson, and Yoshioka (1998) agree that 'the issues of reliability and validity must be dealt with for alternative assessments just as they are for any other type of assessment – in an open, honest, clear, demonstrable, and convincing way' (p. 5). Furthermore, Brown and Hudson (op. cit., p. 657) stress that the term alternative assessments could be harmful because it implies that:

- these assessment procedures are somehow a completely new way of doing things;
- they are somehow completely separate and different; and
- they are somehow exempt from the requirements of responsible test construction and decision making.

Brown and Hudson, thus, propose to call the assessing methods which are commonly known as alternative assessments '*alternatives* in assessment' (emphasis added).

Nonetheless, Lynch (2003) maintains that since traditional testing and alternative assessments are two different paradigms, they require different reliability and validity frameworks. He asserts that within the alternative assessment approach, 'reliability is not necessarily a precondition for validity' as opposed to the traditional testing. Adapted from Meisels, Dorfman and Steele (1995), Hamp-Lyons (1997a) provides a model (below) illustrating the differences between the characteristics of performance/alternative assessment and standardised tests.

Table 2.1: Characteristics of performance assessments and standardized tests (Hamp-Lyons, 1997a, p. 300)

Performance assessment	Standardized test
Criterion referenced	Norm referenced
Contextual objectives	Decontextualized objectives
Modifiable	Uniform
Multidimensional	Restricted dimensions
Longitudinal	Pre/post 'snapshots'
Continuous recording	Discontinuous recording
Monitors progress	Static view of achievement
Extensive behaviour sampling	Restricted behaviour sampling
Reflects quality of work	Reflects speed and accuracy
Promotes student learning	Promotes skill in test-taking
Enhances student motivation	Promotes student anxiety
Instructionally relevant	Instructionally independent
Contributes to classroom change	Imposes institutional change
Informs instructional decisions	Justifies bureaucratic decisions
Useful to parents and others	Unhelpful to parents and others

Though Brown and Hudson (op. cit.) do not agree with the use of the term 'alternative assessment', they recognise that negative washback effects of the assessment on the curriculum could occur when assessment does not correspond to a curriculum's goals and objectives. Positive washback effects could occur when the assessment procedures correspond to the course goals and objectives by using the appropriate assessment format that best matches each objective. Hamp-Lyons (op. cit.), however, stresses that alternative assessment cannot be assumed to have beneficial washback into teaching and learning. Similar to Norris et al. (1998), she asserts that when conducting washback studies of alternative assessment, the researchers must apply the same basis used in traditional forms of assessment (p. 300).

From a different perspective on traditional testing and alternative assessment than described above, Hamp-Lyons (2007b) proposes more fertile directions. In her paper, she argues that there are two different cultures existing in a classroom assessment of English language in ESL/EFL context, 'a learning culture' and 'an

exam culture'. In a learning culture, similar to the underlying concepts of alternative assessment (Alderson & Banerjee, 2001; Lynch 2001a, 2003), 'assessment is shaped by considerations of learning and teaching', whereas 'an exam culture classroom assessment is seen as simply preparation for an externally set and assessed examination' (p. 488). However, she stresses that '[n]either [learning culture or exam culture] is better but they are different' (p. 487; see also Lynch, 2001b; Inbar-Lourie, 2008) and the contrast between learner and exam cultures 'are not static but dynamic and highly contextualized; they are also multi-dimension' (p. 494). Hamp-Lyons also points out three different domains between learning and exam cultures: their focus, their purposes and the voices they ask teachers/educators to listen to (p. 488). Table 2.2 below summarises the contrastive features between the two cultures.

Table 2.2: The two ends of the assessment cultures continuum (Hamp-Lyons, 2007b, p. 494)

Classroom-based assessment	Classical testing
Fluency-focused	Accuracy-focused
Individual-focused	Group- or 'norm'-focused
Achievement/progress focused	Proficiency-focused
Process-focused	Product-focused
Teachers'/student's voices	Rule-makers' voices
Leads to assessment of learning	Leads to 'teaching to the test'

In summary, alternative assessment has become an umbrella term to refer to performance-based assessment as well as the 'alternatives' to traditional discrete-point tests (Fox, 2008). Drawing from the above discussions, great care is needed when implementing alternative assessment, especially on ensuring its quality and impact on learning. It should be noted, however, that in making a decision on which paradigm to adopt, those involved in making the decision, in which classroom teachers must be included, should initially take the purposes of teaching and learning into consideration. Arguably, when the purposes of teaching and learning focus on the construction and administration of standardised or traditional tests in which teaching and testing are separated, the tradition test method is likely to be chosen.

Unfortunately, in this circumstance, the potentials of alternative/ performance-based assessment are neglected. On the other hand, when performance-based assessment is being adopted, teachers are not well prepared to employ them. In this circumstance, the implementation of the performance-based assessment could cause a number of problems among teachers. The present study, as described in the Introduction, was conceived by such circumstance. In the following sections, I present fundamental concepts and empirical studies relating to issues in performance-based language assessment.

2.2 Performance-based Assessment

McNamara (1996) states that a defining characteristic of performance testing is that 'the assessment of the actual performances of relevant tasks are required of candidates, rather than the more abstract demonstration of knowledge, often by means of paper-and-pencil tests' (p. 6; see also McNamara, 1997). Moreover, Davies, Brown, Elder, Hill, Lumley, and McNamara (1999) define performance-based assessment as a 'test in which the ability of candidates to perform particular tasks ... is assessed' (p. 144). Tasks, in the assessment of second language performance, are designed to measure learners' productive language skills through performances which allow learners to exhibit the kinds of language skills that may be required in a real world context (Wigglesworth, 2008, p. 111).

Furthermore, Wigglesworth (ibid.), drawing from McNamara (1996) and Norris, Brown, Hudson and Yoshioka (1998) reports that there are three factors distinguishing performance tests from traditional tests of second language: (1) there is a performance by the candidate; (2) the performance is judged using an agreed set of criteria; and (3) there is a degree of authenticity of the assessment tasks (p. 113). Wigglesworth, based on the same sources, reports that based on the criteria used for judging the performance, there are two types of performance-based assessment. In the first type of performance-based assessment, tasks are used to elicit language to reflect

the kind of real world activities learners will be expected to perform, and in which the focus is on interpreting the learners' ability to perform such tasks in the real world, with language being the means of fulfilling the task requirement rather than an end in itself; McNamara (op. cit.) calls it a 'strong' form of second language performance-based assessment or 'task-based performance assessments' as termed by Norris et al. (op. cit.). In the second type of performance-based assessment, the tasks are used to elicit language samples for the purpose of rating, that is, the focus of the assessment is less on the task and more on the language produced; McNamara (op. cit.) considers it as a 'weak' form of second language performance-based assessment whereas Norris et al. (op. cit.) use the term 'performance based testing'.

Another important characteristic of performance-based assessment discussed by McNamara (1996) is 'a new type of interaction, that between the rater and the scale; this interaction mediates the scoring of the performance' (p. 121). The figure below presents this characteristic of performance-based assessment.

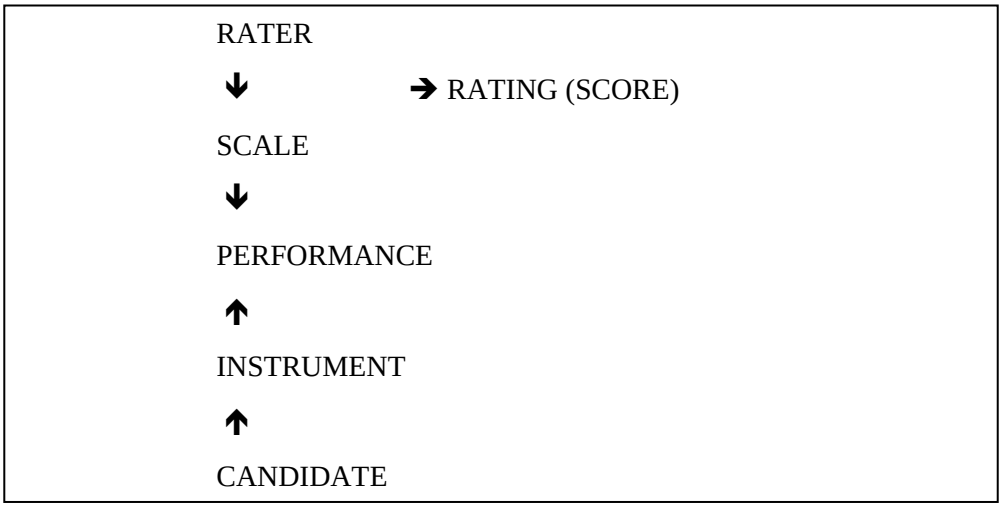


Figure 2.1: Characteristics of performance assessment (McNamara, 1996, p. 120)

In other words, the rater needs to use a rating scale in rating a performance to arrive at a score for that performance. In marking any performance-based assessment tasks, whether in the classroom context or large scale proficiency tests, the markers/raters, or teachers in classrooms, are required to make more complicated judgements than

the right-wrong decisions in multiple-choice, true/false, error-recognition, and other item types where the candidate's responses can be marked as either 'correct' or 'incorrect' (rater issues are discussed in Section 2.4.1). In this type of marking, or sometimes referred to as subjective marking, Alderson, Clapham and Wall (1995) stress that the examiners' job is to assess 'how well a candidate completes a given task', for which they need a 'rating scale' (pp. 106 - 107).

2.3 Rating scales

A rating scale (or proficiency scale) is a 'scale for the description of language proficiency consisting of a series of constructed levels against which a language learner's performance is judged ... The levels or bands are commonly characterised in terms of what subjects can do with the language ... and their mastery of linguistic features' (Davies et al., 1999, p. 153). Rating scales also represent the most 'concrete statement of the construct being measured' (Weigle, 2002). The statements in rating scales are commonly referred to as 'descriptors' which describe 'the level of performance required of candidates at each point on a proficiency scale' (Davies et al., op. cit., p. 43).

It should be noted that in the literature, different terms have been used to refer to a rating scale. For instance, Hudson (2005) reports that sometimes there is a clear distinction between the terms 'rubric' and 'scale' and sometimes they are conflated (p. 207). In this thesis, the term rating scale is used. In addition, because the main focus of the present study is on assessment of written performance (see the introduction of Section 5.2.2), the discussion of rating scales in the chapter is mainly drawn from writing assessment literature. The discussions taken from oral assessment literature are indicated.

According to Alderson (1991), rating scales can be categorised into three types depending on their function and intended audience:

- *User-oriented* scales, with a reporting function, aimed to enable test users – for example, employers and admissions officers – to interpret test results by providing information about typical behaviours of the students at any given level;
- *Assessor-oriented* scales, with a guiding the rating process function, aim to describe guidance for assessors who rate performances by providing typical performances by students at each level;
- *Constructor-oriented* scales, with the function of guiding the construction of tests, aim to provide guidelines for test constructors by providing a set of specifications that students should be able to do at a given level.

In recent language testing and assessment literature, rating scales or scoring methods have been categorised differently by different researchers (e.g. Alderson et al.,1995; Arter & McTighe, 2001; Davies et al., 1999; Hamp-Lyons 1991a; Mertler, 2001; Shaw & Weir, 2007; Weigle, 2002). For instance, Hamp-Lyons (1991a) identifies three types of scoring methods: holistic scoring, primary trait scoring, and multiple trait scoring. (For the discussion of these types of scoring methods, see the following section.) Weigle (2002), on the other hand, identifies three main types of rating scales: primary trait scales, holistic scales, and analytic scales. Weigle does not distinguish multiple-trait scales from analytic scales because she considers that the characteristics of multiple trait scales 'have to do more with procedures for developing and using the scales, rather than with the description of the scales themselves' (p. 109). For the purpose of the present study, I use the terms multiple trait scale and analytic scale interchangeably. In addition, I only explore two types of scales: holistic scales and analytic scales because the teachers, who participated in the present study, had already been familiar with these two terms. In addition, I do not include the primary trait scoring method in the discussion because it is not relevant in the context of the study. This type of scoring method has not been widely used in

second-language assessment (Weigle, ibid., p. 110) but is generally used in research situations particularly in very large-scale data collection (Hamp-Lyons, op. cit.).

2.3.1 Analytic vs. holistic rating scales

With an analytic scale, raters are asked to judge several components of a performance separately as traits, criteria, or dimensions of performance. These components are divided so that they can be judged separately rather than giving a single score for the entire performance (Alderson et al. 1995; Arter & McTighe, 2001; Weigle, 2002). Arter and McTighe (2001) state that analytic scales are used when planning instruction to show relative strengths and weaknesses of a performance, when teaching students the nature of a quality performance, when giving detailed feedback, and when knowing how to precisely describe quality is more important than speed (p. 25). One main advantage of the analytic scoring method over the holistic counterpart is that it provides a higher reliability (Goulden, 1994). Weigle (2002) also agrees that compared to holistic scoring, analytic scoring is more useful in rater training, particularly useful for second-language learners, and more reliable. Moreover, Hamp-Lyons and Kroll (1997) comment that 'a detailed scoring procedure [i.e. multiple trait scoring] requiring the readers to attend to the multidimensionality of ESL writing that may ensure more valid judgement of the mix of strengths and weaknesses often found in ESL writing' (p. 29).

Furthermore, Hamp-Lyons and Kroll (ibid.), drawing from Hamp-Lyons' (1987) study of the scoring procedure for the ELTS (English Language Testing Service, the predecessor of the International English Language Testing Service - IELTS) writing, report that a multiple trait scoring 'helps raters balance their judgments of characteristic ESL features of writing, principally a high frequency of low-order sentence grammar problem, against higher order elements of the writing…' (p. 29). However, Weigle (op. cit.) recognises that the rating time using analytic scoring takes longer than that of holistic scoring because raters need to make more

than one decisions for every script. She also adds that a good deal of the information provided by the analytic scale is lost when scores on the different scales are combined to make a composite score (p. 120).

In contrast, with a holistic scale, raters are asked to give a judgement on a candidate's performance as a whole, or in other words, a single score for an entire performance based on an overall impression of a candidate's work (Alderson et al. 1995; Arter & McTighe, 2001; Weigle, 2002). Thus, the scale used in this method is sometimes called an impression scale. Arter and McTighe (2001) state that holistic scales are used when speed of scoring is more important than knowing precisely how to describe quality, when the performances are simple, and when a quick snapshot of overall achievement is the objective (p. 25). This type of scoring method, nevertheless, has been heavily criticised, especially in the EFL/ESL writing assessment context. Hamp-Lyons (1995, pp. 760-761) points out that:

> a holistic scoring system is a closed system, offering no windows through which teachers can look in and no access points through which researchers can enter. Scores generated holistically cannot be explained to other readers in the same assessment community; diagnostic feedback is out of the question.

Hamp-Lyons' argument is supported by Shi's (2001) empirical study which illustrates that in writing assessment a holistic scoring is not an effective method in distinguishing salient differences of students' performances. From the rater's comments, Shi observes that holistic rating raises questions about the construct validity because the rater's comments demonstrated that they had different understandings of what constitutes good writing.

Furthermore, in the report for the Educational Testing Service (ETS), Hamp-Lyons and Kroll (op. cit., pp. 28-29) point out the inherent nature of holistic scoring being impression marking in a speeded manner. They state that:

> many raters make judgments by responding to the surface of the text and may not reward the strength of ideas and experiences the writer discuss. It is difficult for readers making a single judgment to reach a reasonable balance among all the essential elements of good writing.

Vaughan (1991) employed the think-aloud method to explore what went on in the rater's mind when using a holistic scoring method. One of the main findings she found was that though nine raters had similar training, different raters focus on different elements in the essays and could have individual approaches to reading these essays. Vaughan also found that raters were uncertain whether 'their judgements were within the established criteria' and individual raters relied on his or her own method (p. 121).

Interestingly, Bacha (2001) found high correlations between two sets of scales as well as high inter- and intra-reliability in both holistic and analytic methods. In her study, Bacha had two raters, who were the teachers of the same course, rate 30 essays written by L1 Arabic students of the Freshman English I course using both holistic and analytic scales. The results revealed that the two raters had high inter- and intra-reliability coefficients. However, it should be noted that Bacha reported that a third rater was required in several instances in the data collection when discrepancies exceeded one letter-score range. In addition, only two raters were employed in the study, and it was not mentioned in the study if they were given any training prior to the rating session. This fact could have contributed to the results of the study. In a more recent study, Barkaoui (2007) found similar results to Bacha's (op. cit.) study. In his mixed-method study (Generalizability theory and think-aloud protocol analysis), Barkaoui had four EFL writing teachers rate 32 essays, without any formal training. These essays were written by intermediate EFL university students in Tunisia under exam-like conditions, of which four were used for the think-aloud sessions. Both multiple trait and holistic scales were used. Contrary to the concept that a holistic scale yields lower reliability than multiple trait scale (see Table

2.3 below), Barkaoui found that when the essays were rated holistically, a higher level of score reliability was achieved. He also reported that multiple trait scoring resulted in high rater variability and more ratings were required in order to achieve acceptable dependability indices. However, it should be noted that the scales used were not locally and empirically developed, but from the published scales with minor changes, which might have had an effect on the rating process (for more discussion on empirically derived criteria, see Section 2.3.2 below). The multiple trait scale used in the study was the *Composition Grading Scale*, and the holistic scale was the *EFL Placement Test* developed by Brown and Bailey (1984), and Tyndall and Kenyon (1996), respectively.

In a different context, Iwashita and Grove (2003) studied the assessment of the speaking component of the Occupational English Test (OET) for health professionals in Australia. Iwashita and Grove examined the relationship between analytic and holistic scales used in this testing system where a combined analytic-holistic assessment scale was used. Their study included 13,488 assessments (consisting of assessments by 29 raters) which were collected over eight years. The data was analysed by means of the many-faceted Rasch model programme, FACETS. The results from the analysis of the rating patterns using both analytic and holistic scales suggested that the overall scores did not accurately reflect candidate ability, and the analytic rating could be overrated. Iwashita and Grove concluded that it was possible that using a single holistic criterion may be more accurate and efficient than the combined scale.

Drawing from Bachman and Palmer (1996), Weigle (2002) provides a useful approach to making a decision in choosing between holistic scales and analytic scales in writing assessment. Table 2.3 below presents a comparison of the two types of rating scales based on the six qualities of test usefulness (for more detailed information on test usefulness, see Bachman & Palmer, op. cit.).

Table 2.3: A comparison of holistic and analytic scales on six qualities of test usefulness (Weigle, 2002, p. 121)

Quality	Holistic Scale	Analytic Scale
Reliability	Lower than analytic but still acceptable	Higher than holistic
Construct Validity	Holistic scale assumes that all relevant aspects of writing ability develop at the same rate and can thus be captured in a single score; holistic scores correlate with superficial aspects such as length and handwriting	Analytic scales more appropriate for L2 writers as different aspects of writing ability develop at different rates
Practicality	Relatively fast and easy	Time-consuming; expensive
Impact	Single score may mask an uneven writing profile and may be misleading for placement	More scales provide useful diagnostic information for placement and/or instruction; more useful for rater training
Authenticity	White (1995) argues that reading holistically is a more natural process than reading analytically	Raters may read holistically and adjust analytic scores to match holistic impression
Interactiveness	n/a	n/a

Nevertheless, it should be noted that raters could rate with a 'halo effect' when they employ an analytic rating scale. A halo effect is a rater's failure to discriminate among conceptually distinct and potentially independent aspects of a candidate's performance or a rater's tendency to allow the overall impression of a candidate's performance to influence his or her judgement (Saal et al., 1980; King et al., 1980; cited in Myford & Wolfe, 2003).

2.3.2 Approaches in designing rating scales

After the decision of the type of scale to be adopted, the equally important following step is designing the scales (for the steps of designing rating scales used in the present study, see Appendix A). However, before designing the rating scales, there is another crucial decision to be made, that is choosing a designing approach. From a perspective of designing rating scales in a large-scale testing context, Hudson (2005)

identifies two types of rating scales in relation to criterion-referenced task-based assessment: decontextualised and contextualised. Drawing from Brindley (1998), Hudson describes that the former scale is 'defined independently of content and context… and derived from a theoretical model of language, and attempts to define a decontextualized ability or proficiency' (p. 209); whereas the latter scale 'is behaviourally based and attempts to describe proficiency according to "real-world" performance in specific contexts' (p. 210). Within the behavioural scales, Hudson also identifies two main developmental approaches: intuitive approach (e.g., The Canadian Language Benchmarks, Pawlikowska-Smith, 2000, 2002), and empirical approach (e.g., Common European Framework of Reference for Languages, Council of Europe, 2001; Assessment of Language Performance, Brown, Hudson, Norris, & Bonk, 2002). In this section, I only explore the contextualised approach in developing rating scale because a decontextualised rating scale is not relevant to the present study.

In developing scales in assessing speaking, Luoma (2004, pp. 83 - 86) identifies three methods (within the contextualised approach). The first is 'intuitive methods' in which the development of a scale is based on principled interpretation of experience. The developers, who are usually experienced in teaching and/or material development, may consult existing scales or course syllabus, and then design the scales. The second method is 'qualitative methods'. In this method, the developers ask groups of experts to analyse data related to the scale, which may be the descriptors or samples of performances at different levels. Finally, the third method, 'quantitative method', which mainly addresses scale validation, requires a certain expertise in statistics, such as multidimensional scaling, linear regression, and item response theory. This method is usually carried out by large testing or research institutions.

From another perspective concerning writing assessment, Weigle (2002, pp. 122-124) proposes that once a decision has been made about the kind of rating scale is to be adopted, holistic or analytic, the following factors should be considered:

- Who is going to use the scale?
- What aspects of writing are most important, and how will they be divided up?
- How many points, or scoring levels, will be used?
- How will scores be reported?

After these questions are addressed, the descriptors for levels/bands of the scale can be written. According to Weigle, there are two approaches: *a priori* and empirical. In the *a priori* approach, the 'inherent' ability (for example, a student has ability x) being measured is defined in advanced; whereas in empirical approach, descriptors are derived through the examination of actual performances. Shaw and Weir (2007), in addition, state that the design and development of rating scales for the tests of writing has traditionally relied on *a priori* approach which is based on the experience of an expert and intuitive judgement (p. 162). Nevertheless, they point out that researchers have advocated for more application of the empirically-based approach in developing rating scales. In this approach, samples of actual performances are analysed to construct or re-construct assessment criteria and scales descriptors.

Furthermore, Turner (2000, 2001), Turner and Upshur (2002), and Upshur and Turner (1995) stress the advantageous aspects of empirically derived criteria. Upshur and Turner (1995) strongly believe that scales locally developed by teachers could create positive washback effects on teaching (for fuller discussion on washback, see Section 3.2.1). They point out that because there are no restrictions upon the development of the scale descriptors, the descriptors derived from the interaction among the scale development team reflect instructional objectives. In addition, the development process of the scales and descriptors 'can lead to greater

agreement on the aims of teaching' (p. 11), which can increase the validity of the assessment.

Turner and Upshur (2002) studied the effects of the scale developers and the performance samples on scale content and the scores. The scales developed in the study were based on the empirically derived, binary-choice, boundary-definition (EBB) scale development approach. EBB scales consist of a hierarchical (ordered) set of explicit binary questions relating to the performance being rated. That is, the answer to the first question asked about the performance determines what the next question will be (for more detailed explanations of EBB, see Upshur & Turner, 1995). From the quantitative analysis of the data from the ratings using the empirically derived scale, the findings indicated that the scale development team had a minor effect, whereas the samples used in scale development had a major effect on ratings. It should be noted that the development team were not teachers who were part of a school environment, but they were graduate students who had some experience in the teaching of ESL. Therefore, if the scale development team had actually been the teachers, the findings could have revealed different results. Turner (2000), in addition, investigated the EBB rating scale developed empirically by teachers. Based on qualitative data analysis, she found that scales developed by teachers may have a positive impact on inter-rater reliability when the scale was used within its intended purposes. This is because the teachers brought with them their beliefs, discourse stances, and understanding of curriculum for that particular context. Nonetheless, in these studies, the scales were not intended to be used in a classroom context but for large-scale provincial examinations, though teachers were part of the scale development team in the second study.

Another study on empirically derived scales was carried out by Knoch (2007a). Knoch examined whether an empirically derived scale for writing assessment yielded more reliable and valid ratings than the counterpart *a priori*

derived scale. In this study, the rating scale for the Diagnostic English Language Needs Assessment (DELNA, a test used with first-year international and domestic students admitted to a New Zealand university for diagnostic purpose) was empirically developed and validated from 600 DELNA writing samples. In the validation process, 10 raters, after initial training sessions, rated 100 DELNA writing scripts using the existing DELNA scale, and after 2 months, then used the empirically derived scale. Both scales were analytic. From applying the multifaceted Rasch measurement programme FACETS in the validation process, Knoch found that empirically derived scales resulted in higher inter- and intra-rater reliability than the existing *a priori* scale. She also discovered that the descriptors, which were empirically derived with explicit band level-descriptions, had potential to increase the validity and reliability of a writing assessment because raters could 'count explicit aspects of writing produced by candidates, therefore increasing the chances of raters agreeing on the same score for a script' (p. 23). Nevertheless, Knoch notes that the empirically derived analytic scales operate well when the score for each individual trait was reported separately. Elsewhere, Knoch (2007b) emphasises that 'empirically developed rating scales might lend themselves to being more discriminating and result in higher levels of rater reliability than more conventional rating scales' (p. 122). Furthermore, Knoch (2009) expands Weigle's (2002) classification of rating scales (cf. Table 2.3) by illustrating the differences between the intuitive and empirically developed analytic scales. Table 2.4 below summarises these features.

Table 2.4: A comparison of intuitively developed and empirically developed analytic scales (adapted from Knoch, 2009, p. 299)

Quality	Intuitively developed	Empirically developed
Reliability	Higher than holistic.	Higher than intuitively developed analytic scales.
Construct Validity	Analytic scales more appropriate for L2 writers [than holistic] as different aspects of writing ability develop at different rates. But raters might rate with halo effect.	Higher construct validity as based on real student performance; assumes that different aspects of writing ability develop at different speeds.
Practicality	Time-consuming; expensive.	Time-consuming; most expensive.
Impact	More scales can provide useful diagnostic information for placement, instruction and diagnosis, but might be used holistically by raters; useful for rater training.	Provides even more diagnostic information than intuitively developed analytic scale; especially useful for rater training.
Authenticity	Raters may read holistically and adjust analytic scores to match holistic impression	Raters assess each aspect individually.

Finally, another crucial concept to take into consideration when designing a rating scale in a specific context is 'indigenous assessment criteria'(Jacoby & McNamara, 1999). The concept of indigenous criteria in language education is mainly discussed in the context of English for Specific Purposes (ESP), and is comparatively new and has not been widely investigated. Indigenous assessment criteria refers to the criteria 'used by subject specialists in assessing the communicative performance of apprentices in academic and vocational fields' (Douglas, 2001, p. 175, taking up from Jacoby, 1998). Two significant studies in indigenous assessment include Jacoby and McNamara (1999) and Douglas and Mayers (2000). Jacoby and McNamara (ibid.) compared the findings from two projects the authors were involved in: McNamara and his colleague's studies (McNamara 1996) in the Occupational English Test (OET) in Australia and Jacoby's (1998) doctoral research of conference presentation rehearsals among physicists in

the United States. Douglas and Myers (op. cit.) examined the criteria used by veterinary professionals in performance evaluations of the communication skills of their students in interviewing clients about sick animals.

Both studies illustrate that for specific purposes of language tests in a particular context and when their content and methods are derived from the analysis of the target language use (TLU) situation, the criteria by which the performances are judged should also be derived from the analysis of that particular TLU situation. The main implication from these two studies, for the language testing and assessment in general, is the process in which assessment criteria can be derived. Taking up from the previous discussion on empirically derived rating scale, the scale development team should also incorporate the analysis of TLU in the designing process. In other words, when designing a scale for a particular purpose in a specific context, the team should take into consideration the context for which the scale is to be used; for example, the students, teachers, course syllabus, and so on.

Drawing from the above discussions on different types of rating scales and the designing approaches, in the present study, an analytic rating scale was chosen to be implemented for the written assessment task for the course under investigation. The empirically derived approach was adopted in developing the scale with the needs of the local context taken into consideration.

2.4 Raters

As described in the section above, the rater is one of the most important components of performance-based assessment; therefore, in this section I discuss issues relating to raters. Raters are those who operate:

> a rating scale in the measurement of oral and written proficiency. The reliability of raters depends in part on the quality of their training, the purpose of which is to ensure a high degree of comparability, both inter- and intra-rater. Since raters are human and are therefore subject to individual biases, close attention is paid not only to reliability, but also to analyses of rater bias. (Davies et al., 1999, p. 161)

According to this definition, rating very much depends on the judgement of the raters, given the important roles of raters in performance-based assessment. McNamara (1996) asserts that rater factor is one of the main sources of variability in the scoring of performance-based assessment. He stresses that 'variability associated with ... raters ... is extensive and is a fact of life that must be dealt with ...' (p. 122). Lumley (2002) also illustrates that raters focus on different components of the scale descriptors although they may share similar understandings of the rating criteria.

2.4.1 Rater characteristics and variables

Alderson, Clapham and Wall (1995) emphasise that it is crucial that 'a candidate's score on a test does not depend upon who marked the test, nor upon the consistency of an individual maker'. That is ratings must be reliable. The reliability, for example, of a writing assessment is 'affected by variations in the perceptions and attitudes of those who read the essays, and the kind of training they receive for reading writing assessment' (Hamp-Lyons, 1991a, p. 8). In other words, the reliability of rating has been closely associated with the reliability of raters (Hamp-Lyons, 2007). There are two types of reliability associated with raters: intra- and inter-rater reliability. Intra-rater reliability can be defined as 'the extent to which a particular rater is consistent in using a proficiency scale' (Davies et al., 1999, p. 91); and inter-rater reliability as 'the level of consensus between two or more independent raters in their judgement of candidates' performance' (p. 88).

Drawing from O'Sullivan (2000), Shaw and Weir (2007) present three groups of rater characteristics: physical/physiological, psychological and experiential. They

report that ‘experiential factors’ of rater characteristics have been widely studied by researchers in the field. These factors include education, examination preparedness, examination experience, communication experience, and target language – country residence (p. 168). Furthermore, Reed and Cohen (2001), from reviewing the literature on raters and ratings, summarise four main issues associated with rater characteristics: native/non-native speaker comparisons, raters’ occupations, gender of raters, and personality fit between rater and candidate (pp. 84 – 86). Lumley (2000), from investigating the rating process of the assessment of writing performance, summarises three factors which influence rating process: rater background, rating style and assessment criteria.

Shohamy, Gordon and Kraemer (1992) investigated the effect of raters’ professional background on the reliability of writing assessment. The results, based on Ebel intraclass correlation formula, revealed that the inter-rater reliability coefficients of the raters from different professional background were high, which indicated that trained raters were able to rate reliably regardless of their background. Also concerning professional background of raters, Song and Caruso (1996) investigated two groups of raters: ESL raters and English raters. Based on two-way ANOVA, they found that English and ESL faculty were not significantly different when they scored the essays using analytic scoring (when all analytic features were considered together). However, the English and ESL faculty’s ratings were significantly different when they used a holistic scale. Song and Caruso also found that number of years of teaching experience seemed to affect the way the raters were using holistic scale, but not background and training.

It should be noted that Shohamy et al. and Song and Caruso’s studies employed comparatively less sophisticated statistical tools in their data analysis. Only recently has the multifaceted Rasch measurement and computer software (e.g. FACETS, Linacre, 1989 - 2008) become accessible for language testing and

assessment researchers to investigate aspects (or facets) of performance-based assessment. A. Brown (1995) is among the researchers in language testing and assessment who employed multifaceted Rasch measurement in examining rater facet in performance-based assessment. She investigated the professions (tour guide and language teacher) and linguistic background (Japanese and English native speakers) of raters of the Japanese Test for Tour Guides, which was administered in Australia. She found that these factors did not affect how raters awarded the overall scores when they were given adequate training and explicit assessment criteria. Nevertheless, she reported that the raters differed in the way they applied the scale. Another study supporting Brown's findings is Hill's (1997) investigation into the ratings of Indonesian and English-speaking raters in the English Proficiency Test for Indonesia. Hill, based on a multifaceted Rasch measurement of 13 Indonesian and 10 English-speaking raters, confirmed that the findings did not suggest that native speakers (of English) were more suitable than non-native speaker to rate a test of English language proficiency in this context.

In terms of gender of raters, O'Loughlin (2000, 2002) examined the impact of gender in the IELTS oral interview using discourse analysis of the interview data and the multifaceted Rasch measurement analysis of the scores. The results revealed that the gender of raters (as well as candidates) did not have significant impact on the rating process of the IELTS interview. Nonetheless, O'Loughlin commented that there might be other factors affecting the results of his study, for example, the inherent rating criteria band scale, and the data collection process. He admitted that 'gendered differences are not inevitable in the testing context', and 'gender competes with other aspects of an individual's social identity in a fluid and dynamic fashion [in particular contexts]' (O'Loughlin, 2002, p. 190).

Furthermore, Lumley (2000) employed multifaceted Rasch measurement as well as other quantitative and qualitative methods to investigate the rating process of

the Special Test of English Proficiency (*step*), a high-stakes test administered on behalf of the Australian government as part of the immigration process. In the study, Lumley examined the rating process of four accredited *step* raters, who were from similar backgrounds. Lumley (2002) reports that the central feature of performance-based assessment is the rater, not the scale, because the rater is the person who uses the scale to make decisions. From the analysis of the qualitative data, he observes that it is the rater who decides - which features of the scale to pay attention to; how to arbitrate between the inevitable conflicts in scale wordings; and how to justify her impression of the text in terms of the institutional requirements represented by the scale and rater training (p. 267). Lumley (ibid.), in addition, indicates that rating scales are only 'tools' for raters to use when they read texts, and not necessarily a valid statement of how they actually apply the scales because of their limited ability to describe texts adequately. As the scales do not include all eventualities, raters have to develop their own strategies to help them deal with problematic aspects of the rating process (Lumley, 2000, p. 310).

Based on the above discussions, when recruiting the participants in the present study, I tried to involve participants, who were teachers and raters, from different backgrounds - for example, their educational background, gender, and experiences. In addition, the multifaceted Rasch measurement, with the aid of FACETS, was used in the preparation stage for the main study to investigate teachers' behaviours in rating.

2.4.2 Rater training

Alderson, Clapham and Wall (1995) point out that one of the most important issues to consider in teacher assessment is rater monitoring. Alderson et al. state that training the examiners or raters could provide them with 'competence and confidence' (p. 128). In addition, they stress that it is the responsibility of the institution to ensure examiners to mark the test as reliably as possible by designing appropriate quality

control procedure. Quality control procedure, they argue, can ensure the intra-rater and inter-rater reliability of the assessment. Likewise, in order to improve the quality of rater-mediated assessment, McNamara (2000) emphasises the moderating meeting scheme providing initial and ongoing training to raters. Alderson et al. (op. cit.) also add that on a regular basis, tests should be routinely monitored, after each administration item and subtest analyses and descriptive statistic analyses should be conducted, raters should be monitored, and post-test reports should contain information for any future modification. In the same vein, Davies et al. (1999, p. 161) state that the reliability of raters depends, partially, on the quality of their training, which aims to ensure a high degree of both inter- and intra-rater. In addition, Lumley (2002) stresses that rater training and reorientation allows raters to 'learn or (re)develop a sense of what the institutionally sanctioned interpretations are of task requirements and scale features, and how others related personal impressions of text quality to the rating scale provided', which increase the reliability of rating (p. 267). Shohamy, Gordon and Kraemer (1992) found that intensive procedural training could improve inter- and intra-rater reliability. In their study (as described in Section 2.4.1 above), they discovered that the scores of the professional English teachers, who received training, were stable after a three weeks interval.

It is, however, important to be aware that training on its own cannot guarantee that raters will mark as they are supposed to (Alderson et al., op. cit., p. 128). In addition, Hamp-Lyons (2007b) states that rater training can influence how teachers judge their students' language performances, but making judgements still remains subjective because it is based on individual teacher's experiences. Davies et al. (op. cit., p. 161) support that:

> rater training shows that training reduces extreme differences in severity between raters and makes raters more internally self-consistent, but that significant differences in severity between raters remain; further, that rater characteristics (relative severity, self-consistency) vary over time.

Vaughan (1991) also reports that although the raters in her study had similar training, when rating essays using holistic scales different raters focused on different elements of the scales and could have individual approaches to reading essays.

Weigle (1994) investigated the effects of training on raters of ESL compositions using both quantitative and qualitative methods. In this study, Weigle included 16 raters, of which half were inexperienced raters (who were the focus of the study). The data was collected before, during and after training sessions. The data revealed that the training helped the inexperienced raters to understand and apply the rating criteria. The training also brought these raters 'more or less in line with the rest of the raters' (ibid., p. 214). However, a new insight was revealed when Weigle later applied the multifaceted Rasch measurement to analyse the data. From the analysis, Weigle (1998) found that 'rater training cannot make raters into duplicates of each other, but it can make raters more self-consistent' (p. 281).

Lumley and McNamara (1995) also report that the results of rater training are not long lasting. Lumley and McNamara compared the test scores from the Occupational English Test administered in Australia which obtained from two rater training sessions, 18 months apart, and a subsequent operational administered of the test (about two months after the second training session). They employed the multifaceted Rasch measurement and found the inconsistencies and changes of raters' behaviours between the rater training sessions and the actual test administration, especially from the second training session and the operational administration. Lumley and McNamara, thus, suggested that rater training should be conducted at every administration of the test.

Different from face-to-face rater training, Elder, Barkhuizen, Knoch, and von Randow (2007) explored online rater self-training of 8 ESL raters rating the DELNA (Diagnostic English Language Needs Assessment) test administered at a university in New Zealand (for the description of the rating scale used in the study, see Knoch,

2007a, Section 2.3.2). Elder et al., similarly to Lumley and McNamara (op. cit.), employed the multifaceted Rasch measurement in investigating the consistency of rater characteristics over time after the trainings. It should be noted that the online training was not to replace the traditional training, but an online package, with 25 benchmark writing samples, was for raters to retrain themselves before the actual marking. From the questionnaires and test score data collected before, during and after each of the online rater training programme, Elder et al. found that the online training programme had minimal impact on the overall reliability of the ratings. The programme did not increase the intra-reliability, nor did it decrease the individual biases of the raters in relation to different dimensions of the rating scale.

After this study, Knoch, Read and von Randow (2007) did a further investigation of this online rater training by comparing it with a traditional face-to-face training. In this study, eight raters received online training and another eight face-to-face training. The test scores, questionnaires and interviews were collected in four phases: pre-training rating, training, post-training rating, and post-training feedback. With the aid of the computer programme FACETS, Knoch et al. found that both forms of trainings were effective in increasing inter-rater reliability. However, online training might be more successful in decreasing differences between raters in terms of harshness and leniency, whereas face-to-face training might be more successful in reducing the halo effect. The halo effect occurs when a rater awards the same score, based on his/her overall impression, for all categories in an analytic rating scale. In other words, the raters do not use analytic scales in an analytical manner. Nonetheless, from the analysis of qualitative data, Knoch et al. found that some raters seemed to prefer a mixture of the two methods.

Though I previously discussed some drawbacks of rating training, it is one of the most crucial procedures in ensuring the quality of rating process cycle. Therefore, in this section, I explore the recommended rater training procedures.

A rating training prepares raters for the task of judging candidate performances. It mainly involved in the process of the raters familiarised with the test format, test tasks, rating criteria, and exemplar performances at each criterion level (Davies et al., 1999, p. 161). Building on White (1984), Weigle (2002, pp. 130 - 131) sets up a guideline for training raters of writing assessment. The first step, the leader (or preferably a team) should read through the scripts to find anchor/benchmark scripts that exemplify the different bands/levels on the rating scale. The scripts that exemplify certain problematic situations should be included. After that, the first set of scripts is generally given to raters in order (e.g. from highest to lowest) with the appropriate scores indicated. Nonetheless, the purpose of this activity is to familiarise raters with the scale and illustrate certain features of the rating criteria. When the raters are comfortable with the scale, a set of scripts, including one script at each level in random order, should then be given. Finally, raters should work with more problematic sets of scripts, which may have more than one script at a given level, or, may be less clearly representative of certain points of the scale. Furthermore, Weigle recognises that it is important to note that getting a large group of raters to agree on exact scores is virtually impossible, and some disagreement among raters is expected. Thus, it is crucial to inform the raters that they are not required to be perfectly accurate all the time. However, the raters who consistently rate lower or higher than the rest of the group should be given feedback and perhaps retrained.

However, Alderson, Clapham and Wall (1995) have a rather different view of how to conduct rater training or ‘standardisation meetings’. While Weigle (op. cit.) suggests that the consensus scripts should be given with the scores indicated, Alderson et al. state that the raters should not be shown the decisions made by the committee ‘to prevent examiners from being influenced by the original committee’s reasoning before they have had a chance to try out the scale and think for themselves’ (p. 112). The consensus scripts are those scripts that represent ‘adequate’ and ‘inadequate’ performances, and scripts which present common problems raters often

face but are rarely described in rating scales. The raters should try out the rating scale on the consensus scripts which are given before the meeting.

The first stage of the meeting should be devoted to discussing the consensus scripts to find out if all raters agree on the marks they have given, and to work out why they have had problems if they do not agree. The aim of this activity is to help all raters to match the marks of the original committee. Thus, the committee's consensus scores should not be indicated on the scripts. After that, the problematic scripts should be presented, together with guidelines on what raters should do in these cases. Then, further practice in marking should be provided with another set of scripts. It should be noted that for Alderson et al., if disagreements among raters were from unclear wording or concepts in the rating scale, the scale should be edited. After the scale is edited, it should be given to the raters who will proceed to rating candidate's performances. Alderson et al. emphasises that after this point 'no further changes should be made to this scale' (p. 113). Similar to Alderson et al., McNamara (2000) states that the rater rating or moderating meeting scheme is a process which involves individual raters independently marking a series of different levels of performance. Then in groups they have to share their marks with other raters. The differences are noted and discussed in detail by referring to the interpretation of different levels of descriptors of individual raters. The purpose of the meeting is to try to bring about a general agreement on the relevant descriptors and rating categories. Because in the present study, revising the scale was part of the rater training, I follow the guidelines of Alderson et al. (1995) and McNamara (2000).

2.5 Conclusion

In this chapter I have explored in detail the concepts and empirical studies pertaining to performance-based assessment focusing on rating issues. The discussion focuses on rating scales and raters. In the section on rating scales, I included studies of types of rating scales and approaches in designing rating scales. Studies in rater

characteristics and variables and rater training are described in the raters section. From reviewing the literature in performance-based assessment, I have become aware of the significant roles of how rating scales and raters play significant roles in ensuring the quality of the assessment. In the present study, the development of an empirically derived rating scale was utilised in providing an in-service training for teachers who are raters of their students' performances.

3 Teacher Change and Professional Development Programme in Language Assessment

In language education, the main focus of studies in teacher change has been on, for instance, practical experiences for curriculum and materials development, classroom-centred or teacher research, teacher cognition, and innovation and teacher change (Crandall, 2002). In this chapter, I firstly introduce the studies in teacher change from both general education and language education perspectives. Secondly, I discuss the concept between teacher change and language assessment, including the definitions of related terms and empirical studies in washback and impact of assessment. Furthermore, I include the discussions on teacher change and professional development (PD) in general education and language testing and assessment. Finally, I conclude this chapter with the different views from the innovation theory.

3.1 Teacher Change

The nature of change is multifaceted and complex (Richards, Gallo & Renandya, 2001). From reviewing related literature, Richards et al. (ibid.) states that change can refer to many things, for instance, attitudes, beliefs, knowledge, understanding, self-awareness, and teaching practices. They also report the assumptions below about the nature of teacher change considering underlying current approaches to teacher PD:

> Teachers' beliefs play a central role in the process of teacher development; changes in teachers' practices are the result of changes in teachers' beliefs; and the notion of teacher change is multidimensional and is triggered both by personal factors as well as by the professional contexts in which teachers work. (p. 41)

Furthermore, Tsui (2007) summarises four major factors shaping teachers' conceptions of teaching and learning: personal background and life experiences; their disciplinary training; their teaching and learning experiences; and their professional training. She adds that these conceptions may change or be modified when teachers gain experience or as they encounter critical incidents, and/or they may be very resistant to change (p. 1055). Clarke and Hollingsworth (2002, p. 948) review studies in teacher professional growth and describe alternative perspectives on teacher change as follows:

- Change as training – change is something that is done to teachers; that is, teachers are 'changed'.
- Change as adaptation – teachers 'change' in response to something; they adapt their practices to changed conditions.
- Change as personal development – teacher 'seek change' in an attempt to improve their performance or develop additional skills or strategies.
- Change as local reform – teachers 'change something' for reasons of personal growth.
- Change as systemic restructuring – teachers enact the 'change policies' of the system.
- Change as growth or learning – teachers 'change inevitably through professional activity'; teachers are themselves learners who work in a learning community.

Sakui and Gaies (2003), from reviewing studies in the field of applied linguistics, have found that studies on teacher change have focused, for example, on teachers' beliefs and behaviours on the use of written language in beginners' classrooms, teacher beliefs in reading instruction, in grammar teaching, in communicative language teaching, and teachers' perceptions of innovations. Drawing on work by Breen (1991), Burns (1992) adds that the study of change should involve 'the challenging and questioning of one's beliefs' in addition to the perspectives and

reflections of teachers themselves (p. 64). Burns also asserts that to change and enhance the teaching and learning of language, it is crucial to explore both theory and practice. In her view, 'theory is what researchers and textbooks writers "do" while practice is the real stuff of daily classroom life' (ibid.). Kagan (1992), based on numerous studies in teacher beliefs, points out that greater attention to the social and institutional contexts of classrooms is required in studies of what language teachers do. She also proposed that further research into the process of transformation of language teachers' cognitions and practices as they accumulate experience is required, in addition to the study of cognitions and their patterns amongst groups of teachers working in a similar context. From Borg's (2003) review, however, it should be noted that 'behaviour change does not imply cognitive change, and the latter does not guarantee changes in behaviour either' (p. 91).

Similarly, Freeman (1989) describes language teacher education as an interactive process between the teacher and the collaborator (for example, teacher educator, trainer, supervisor, or colleague). The two individuals engage in a process 'to generate change in some aspects of the teacher's decision making based on knowledge, skills, attitude, and awareness' (p. 38). Freeman (ibid.) also points out four characteristics of change:

- Change does not necessarily mean doing something differently; it can mean a change in awareness.
- Change is not necessarily immediate or complete.
- Some changes, for example the number of techniques used to correct, are quantifiable; whereas other changes, for example a change in attitude, are not.
- Some types of change can come to closure and others are open-ended.

Brindley (2008, p. 370) reports the following key messages which emerge from the studies in curriculum and assessment reform in language teaching contexts and that are reflected in the mainstream educational literature:

- centrally driven educational reform initiatives rarely succeed. The changes that last are generally those that are local and locally adapted;
- successful change involves shared control and decision-making;
- teachers are the key factor in the implementation of reform; the likelihood of whether a change will be implemented depends on the degree to which it is linked to daily classroom practice; and
- ongoing in-service education is vital in ensuring the sustainability of an innovation.

Another domain of language teacher change which has been examined and fairly well established in the field of applied linguistics is language teacher cognition (Borg, 2003, 2006). The following section provides a brief overview of this domain of inquiry relevant to the present study. Borg (2003, 2006) reviews more than 180 studies in teacher cognition in the areas of first, second and foreign language contexts published between 1976 and 2006. He has found that research in teacher education can benefit greatly from focusing on the content, structure, and development process in language teachers' cognition. He points out that the studies of teachers' cognition include:

> what teachers at any stage of their careers think, know or believe in relation to any aspect of their work, and which, additionally but not necessarily, also entail the study of actual classroom practices and of the relationships between cognitions and these practices. (Borg 2006, p. 50)

In other words, teacher' cognitions include teachers' beliefs, knowledge, attitudes and practices.

However, it should be noted that there is a multiplicity of concepts and labels adopted in the teacher cognition research. Borg (ibid., pp. 47 – 49) has compiled the terminology in this research area; for example, BAK, beliefs, epistemological beliefs, conceptions of practice, knowledge about language, practical knowledge, personal

practical knowledge, maxims, pedagogical reasoning perception, theories for practice. It should be noted that these terms, especially *beliefs*, *knowledge* and *attitudes*, have been used and defined by different researchers with different meanings and the same constructs have been termed differently. In this thesis, I adopt the definitions of the terms 'belief' and 'attitude' proposed by Dörnyei (2005). Dörnyei (ibid., p. 214) defines attitudes to have 'a stronger factual support [than beliefs]' whereas beliefs 'are more deeply embedded in our mind and can be rooted back in our past or in the influence of the modelling example of some significant person around us'. In addition, Pajares (1992) has made a clear distinction between beliefs and knowledge. Pajares (ibid., p. 313) states that beliefs are 'based on evaluation and judgement' and knowledge is 'based on objective fact'. He also adds that beliefs are 'an individual's judgement of the truth or falsify of a proposition' (p. 316).

Johnson (1994, p. 439), from reviewing extensive studies, summarises the following basic assumptions on teachers' beliefs:

- teachers' beliefs influence both perception and judgement which, in turn, affects what teachers say and do in classrooms;
- teachers' beliefs play a critical role in how teachers learn to teach, that is, how they interpret new information about learning and teaching and how that information is translated into classroom practices; and
- understanding teachers' beliefs is essential to improving teaching practices and professional teacher preparation programmes.

Johnson (ibid., p. 440) also stresses that in investigating into teacher's beliefs, it is crucial to infer beliefs from the statements that teachers make about their beliefs as well as examine teachers' intentions and what they actually do. In her study, Johnson examined pre-service teachers' beliefs during a practicum teaching experience of four students enrolled in an MA programme in Teaching English as a Second Language. Johnson examined the narratives, intentions, and instructional practices of these pre-

service teachers and found that these teachers questioned their own beliefs, recognised the inconsistencies in their own practices, and were seeking to project images of themselves as teachers and of teaching through a process of becoming reflective and conscious of their own practices (p. 450). In other words, critiquing one's own beliefs and practices is a crucial part of professional learning in a pre-service training.

In a context of cognition change during in-service training, Freeman (1993) investigated changes in practice and thinking of four high school French and Spanish teachers doing an in-service MA programme in teaching in the United States. In this longitudinal qualitative study, Freeman points out four main concepts that emerged from the data: conceptions of practice, tensions of these concepts, the process of articulation, and local and professional language. In terms of teachers' conceptions of practice, Freeman found that when these teachers entered new situations, they brought with them conceptions of teaching which were not explicitly articulated. He stresses that these conceptions then surfaced as tensions in the in-service programme. These tensions were 'expressed as discomforts or confusions which interfere with the teachers' translating intention into action in the classroom' (p. 488). Consequently, Freeman proposes that it is important for teachers to recognise these tensions in order to develop their classroom practice. For process of articulation, Freeman found that the teachers in his study did not have an opportunity to talk about their thinking and classroom practice; the process which would enable them to critique their classroom practice. The data also revealed that in order to effectively critique their practice, teachers had to 'combine the new professional knowledge [professional discourse of education] of the teacher education program with their local language explanations [the vehicle through which teachers explain what goes on in their teaching]' (p. 489). Finally, Freeman (p. 495) concludes that 'the notion of [teacher] change becomes more complicated because it is no longer possible to simply use behavior as the criterion by which to access it'.

Furthermore, also in an in-service context, Woods (1996) points out that the constructs of beliefs, assumptions and knowledge (BAK) are interwoven and integrated. In other words, they are points on a spectrum of meaning rather than being distinct concepts (Borg, 2006, p. 92). Woods (op. cit.) stresses that these constructs affect the decisions a teacher makes in interpreting events related to teaching. From his qualitative data collected from eight teachers in Canada, he concludes that 'BAK develops through a teacher's experiences as a learner and a teacher, evolving in the face of conflicts and inconsistencies, and gaining depth and breadth as varied events are interpreted and reflected upon' (p. 212). Burns (1996) also investigated how six experienced ESL teachers' thinking and beliefs inform their planning, decision making and curriculum enactment. She found that the thinking and beliefs the teachers in her study brought with them into classroom processes appeared to be 'highly significant but are frequently unconscious and implicit' (p. 175). Nonetheless, Burns points out that these thinking and beliefs appeared to 'activate and shape patterns of classroom interaction, roles and relationships and, therefore, to create for learners particular kinds of opportunities for learning' (ibid.). Finally, Burns puts forward that the investigation into teachers' thinking and beliefs would offer 'critical insights into the nature of professional growth and the forms of in-service and professional development support which would most appropriately enhance [classroom work]' (p. 176).

The focus of the investigation on teacher change in the present study, therefore, became the beliefs, knowledge, attitude, understanding, self-awareness, and practices of teachers. Teachers may change as a result from being been involved in training or professional development activities. In addition, they may change because the local community changes. However, teachers may not at all change or resist change.

3.2 Assessment and Teacher Change

In general education, it has been noted that classroom assessment is a complex operation especially in contexts where there are frequent changes in the assessment systems, which can cause a great deal of confusion and anxiety among teachers (Mavrommatis, 1997). In addition, there is a great deal of variability in assessment practices among teachers. According to a survey conducted by Cizek, Shawn and Fitzgerald (1995), teachers' assessment practices are highly variable and unpredictable depending on their characteristics such as gender and years of experience. Since the practices of teachers in classroom assessment are varied, there have been attempts to set standards for teachers in terms of their competence in educational assessment by, for example, the American Federation of Teachers, National Council on Measurement in Education and the National Education Association (1990, see Appendix B). However, Cizek et al. (op. cit.) report from their review of the literature that 'teacher's assessment practices do not necessarily conform to what measurement specialists would consider to be sound testing and grading practice' (p. 173).

Recently, assessment done by teachers in a classroom context has been one of the central interests in language education. Teachers have to teach and assess students, especially in ESL/EFL contexts in which performance-based assessment is implemented. In these contexts, many methods are used to collect information about the abilities of the students apart from traditional paper-and-pencil tests. For instance, Genesee and Upshur (1996, p. 4) propose 'evaluating without tests' which include observations in the classroom, portfolios, conferences, journals, questionnaires, and interviews (see also the discussion on alternative assessment, Section 2.1). However, it is well known that for language teachers, testing and assessment are considered as 'the somewhat arcane province of "expert" and of marginal relevance to everyday classroom concerns' (Brindley, 2001, p. 127). Leung (2004) stresses that even

teachers working within the same curriculum and assessment framework might have different practices in assessment. In addition, Davison (2004) reports on the results of her study on large-scale criterion-referenced assessment in schools in Australia that there is 'a great diversity in teachers' approaches to assessment, influenced by the teachers' prior experiences and professional development, by the assessment frameworks and scales they used, and by the reporting requirements placed on them by schools and systems' (p. 39).

3.2.1 Impact of assessment: definitions and related concepts

It is well accepted that 'assessments come in all shapes and sizes, ranging from international monitoring exercises to work with individual pupils in the classroom. These assessments each have their purposes and their *consequences'* (Stobart, 2003, p. 139, emphasis added). Assessment, thus, has been viewed as a powerful tool and used by 'authorities' to create change (Shohamy, 2007; see also Shohamy, 2001; McNamara, 2008). The consequences or effects of assessment are known by language educators as 'washback (or backwash)' and 'impact'. Bachman and Palmer (1996) acknowledge that tests have an impact 'on society and educational systems and upon the individuals within those systems' (p. 29). In general education, Cheng (2008) and Hamp-Lyons (1997a) report that the concepts of these two terms have been well documented, but referred to differently. For instance, 'measurement-driven instruction', which implies that testing should drive teaching and learning; 'curricular alignment' which focuses on the relationship between test content and curriculum and teachers' training practices; and 'consequences' which focuses on the intended or unintended and positive or negative aspects of high-stakes testing on instruction, students, teacher and the school. However, it should be noted that in general education, the term 'washback' is not used, but 'impact' is used to refer to the effects of high-stake tests (Hamp-Lyons, ibid., p. 297). Wall (1997) points out that washback

is sometimes used interchangeable with ‘impact’, but the term ‘washback’ is ‘more frequently used to refer to the *effects* of tests on teaching and learning’ (p. 291, emphasis added). These effects are usually perceived as being negative because teachers could be forced to do what they ‘do not necessarily wish to do’ (Alderson & Banerjee, 2001). Alderson and Banerjee (ibid.) also report that researchers have argued that ‘tests are potentially also “levers for change” in language education… [i.e.] good tests should or could have positive washback’ (p. 214). In other words, the effects of tests could be either positive or negative. Similarly, Alderson and Wall (1993) comment that when conducting studies in washback, researchers need to consider both negative and positive effects because washback entails that the effects of tests can be either intended or unintended, and directly or indirectly.

However, Alderson and Banerjee (2001), Cheng and Curtis (2004) and Turner (2006) do not distinguish the terms ‘impact’ and ‘washback’. On the other hand, Wall (1997) distinguishes impact from washback and defines impact as ‘any of the effects that a test may have on individuals, policies or practices, within the classroom, the school, the educational system or society as a whole’ (p. 291). In addition, Hamp-Lyons (1997b) stresses that washback is one form of impact, and language testers must view impact as ‘pervading every aspect of … [the] instruments and scoring procedures’ (p. 299). In other words, the term ‘washback’ is too narrow whereas ‘impact’ includes the effects beyond the classroom. Rea-Dickins and Scott (2007b), in contrast, argue that ‘Rather than simply being an aspect of “impact”, washback perhaps follows from impact, equally unpredictable and changeable, but not necessarily malleable by external agency’ (p. 5). Bachman and Palmer (1996), though they do not explicitly distinguish the two terms, point out that the impact of test use operates at two levels:

- a micro level, in terms of individuals who are affected by the particular test use, especially, test takers and teachers; and

- a macro level, in terms of society and education systems.

In this thesis, I follow Wall's (op. cit.) and Hamp-Lyons' (op. cit.) proposition that the review of washback studies in the following sections is part of the all-encompassing of impact studies. However, the word 'impact' may be used in a non-technical sense to refer to 'effect'.

Drawing from Alderson and Wall's (1993) washback hypotheses and Hughes (1993), Bailey (1996) proposes a simple washback model (see Appendix C). In the model, Bailey argues that not only tests have impact on participants (including students, teachers, materials writers and curriculum designers, and researchers), the products (including learning, teaching, new materials and new curricula, and research results), but participants may also have impact on the tests. This impact is what van Lier (1989) calls 'washforward' (cited in Bailey, 1996). Based on a review of major washback studies, Wall (2000) summarises the factors which account for test impact. Her list includes: teacher ability, teacher understanding of the test and approach it was based on, classroom conditions, lack of resources, management practices within the school, the status of the subject in the curriculum, feedback mechanisms between the testing agency and the schools, teacher style, commitment and willingness to innovate, teacher background, the general social and political context, the amount of time that has passed since the introduction of the exam, and the role of publishers in materials design and teacher training (p. 502).

Having realised the impact of language tests, Hamp-Lyons (2002) emphasises the ethical responsibilities language testers have to take when designing or administering, and scoring a writing test as well as taking and utilising test scores. She recognises this process as a form of 'social engineering' which could be 'beneficial and dangerous' (p. 13). She concludes that '[a]ccepting a shared responsibility for the impact of writing assessment practices will put consideration of our own ethical behaviour at the top of our agenda' (p. 14). Hamp-Lyons (1997a)

points out that because tests have impact on test takers, classroom, school systems and even whole society, therefore, testers should 'avoid negative impact and maximize the possibility of positive washback' by taking 'account of impact, and work consciously in test development, administration, reporting, and *advertising*' (p. 326, emphasis in original). In other words, language testers have to accept responsibility for all those consequences they are aware of (Hamp-Lyons, 1997b, p. 302). Stobart (2003) also agrees that assessment is not a neutral process, thus, always has consequences. Therefore, the task of educators and language testers is to make sure that the assessment is as constructive as possible, especially for the candidates (ibid., p. 140). Moreover, Hamp-Lyons (2001, p. 227) urges language testers to:

> critique everything we do, and to take that critique onward and look at the impact we have on test takers, other stake holder groups, and on society, and we must not flinch from accepting some responsibility for the uses made of the tests we have been involved in ...

Nonetheless, Davies et al. (1999) argue that 'language test developers cannot, of course, be held responsible for uses of their tests which are beyond their control' (p. 31). The pertinent question left unanswered is, then, 'where and when we [language testers] decide to let our responsibility drop?' (Hamp-Lyons, 2001, p. 227). For further discussions on ethical issues (including fairness and bias) in language testing and assessment, see the special issue of *Language Testing* (edited by Davies, 1997, Vol. 14, No.3) and *Fairness and validation in language assessment: selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (edited by Kunnan, 2000).

3.2.2 Empirical studies

In the early 1990s, there were not many studies in washback carried out in the field of language testing and assessment. Alderson and Wall (1993) were among the first

researchers to call for more studies to be carried out in this area. Wall and Alderson (1993) examined the effects of the new O-level examination (leaving school exam) on English teaching in secondary schools in Sri Lanka. The study was carried out over the period of three years (including a baseline study) and over five different areas of the country. Drawing from Alderson and Wall's review (op. cit.), Wall and Alderson stressed that, by the time it was published, this study was the only study that included classroom observation as one of its research methods. From the analysis, they found that there was evidence of positive and negative washback on the content of teaching, there was no evidence of washback on teaching methodology, and there was evidence of positive and negative washback on the way teachers and local education offices designed tests. In other words, the introduction of the new examination had impact on '*what* teachers teach but not on *how* they teach' (p. 68, emphasis in original).

In a different context to Wall and Alderson's (ibid.) study, Watanabe (1996) investigated whether the use of grammar translation in classrooms was in fact due to grammar translation used in university entrance examinations in Japan. Watanabe compared two teachers who taught at private extracurricular institutions preparing for the university entrance exams using classroom observation (two different exam preparation courses) and interview methods. The findings revealed that the presence of translation questions did not affect these two teachers in the same way, that is, translation-oriented entrance exams had washback effects on some teachers, but not on others (p. 330).

Alderson and Hamp-Lyons (1996) used similar research methods as Watanabe to examine the washback effects of the large-scale high-stakes proficiency test TOEFL (Test of English as a Foreign Language) on preparation classrooms at a language institute in the United States. In their studies, Alderson and Hamp-Lyons interviewed (individual and group) and observed two teachers (field notes and audio-

recording in TOEFL preparation courses and non-TOEFL courses), and interviewed (group) three sets of students. Different from Wall and Alderson's (op. cit.) study in Sri Lanka, Alderson and Hamp-Lyons found that 'the TOEFL affects both *what* and *how* teachers teach, but the effect is not the same in degree or in kind from teacher to teacher' and, 'the simple difference of TOEFL versus non-TOEFL teaching does not explain *why* they teach the way they do' (p. 295, emphases in original). They finally suggested that the amount and type of washback would depend on: the status of the test (the level of the stakes); the extent to which the test is counter to current practice; the extent to which teachers and textbooks writers think about appropriate methods for test preparation; and the extent to which teachers and textbook writers are willing and able to innovate (p. 296). Alderson (2004) reflects on this study and notes, 'it is at least as much the teacher who brings about washback, be it positive or negative, as it is the test' (p. x).

Cheng (2005, see also 1997, 1998, 1999) investigated the washback of the Hong Kong Certificate of Educational Examination in English (HKCEE), a high-stakes public exam, in secondary schools. In this study, Cheng employed multiple methods, quantitative and qualitative methods, to explore the washback effect at the macro and micro levels. At the macro level, perspectives from different stakeholders were analysed, and at the micro level, the washback on teachers, students, and classroom were scrutinised. Similar to Wall and Alderson (op. cit.), the findings revealed that the introduction of the new examination affected what teachers teach, but not how. In other words, the change of the examination could change teachers' classroom activities, but it did not change teachers' beliefs and attitudes about teaching and learning, the roles of teachers and students, and how teaching and learning should be carried out.

Wall (2005; see also 1996; 2000) revisited the Sri Lanka impact study (Wall and Alderson, 1993, see the above discussion) using the insights from educational

innovation theory. In the analysis of the data, Wall vigorously applied Henrichsen's (1989) hybrid model of the diffusion/implementation process of innovations in English language teaching to understand the antecedents, process and consequences of the impact of the new national examinations on classroom teaching (for a fuller discussion of the model, see Section 2.5.4). Wall concludes that the impact of exams is complex, which 'should not be seen as a natural or inevitable consequences of introducing a new examination into an educational setting' though 'the design of the examination will always have some effect on the way that teachers react to it' (p. 279). There are, nevertheless, many factors (especially those described in Henrichsens' model) determining the impact of the exam on individual teachers; for example, the teachers' view of the impact of the exams on their teaching, context before the introduction of the exam, characteristics of the textbook series, characteristics of the exams, characteristics of the system (e.g. classroom factors, educational administration, political factors), and characteristics of the users (i.e. teachers and students). Finally, Wall proposes that it is very valuable to use a framework from educational innovation theory, such as Henrichsen's, in examining the impact of examination projects, especially of changes.

Turner (2006), in a different context, conducted a survey to examine the impact of provincial exams in Quebec, Canada, on 153 ESL school teachers across the region. The study focused on the views of the teachers concerning the changing of the educational testing system and the consequences of the changes on their behaviours and classroom activities. The Quebec education system was changing toward a more school-based assessment with emphasis on speaking (for more detail on school-based assessment in Hong Kong context, see Davison, 2007). From the questionnaire survey, the data revealed that 'teachers may or may not embrace the changes, but they cope with them as part of their work and integrated them into their teaching practice' (p. 71). Turner also discovered that teachers, in this context, wanted to take part in the change process. Nevertheless, the results showed no

evidence of the influence of the provincial exam on teachers' views of their behaviours. It should be noted that this study did not involve any qualitative data. Interviews with teachers might have shed light on the reasons underlying the teachers' enthusiasm in participating in the curriculum changes and their perceptions of the impact of the exam.

Wall and Horák (2006; 2008) investigated the impact of the changes of the TOEFL test (to the Internet-based test, iBT) on teaching and learning in preparing students to take the test from a teachers' point of view in Central and Eastern Europe. Wall and Horák interviewed six teachers twice a month over the period of five months and found that at the beginning of the study, teachers' awareness of the changes in the TOEFL was quite low but grew during the study. They also found that the teachers had a positive attitude toward the introduction of speaking test and the integrated writing task. The teachers also expressed that the changes of the test would result in changes in their classroom. Finally, Wall and Horák assert that the availability and quality of the information about the test and test preparation materials would be a major source contributing to teachers' reaction to the changes and how they would cope with the changes.

3.2.3 Summary

From the above discussions, it can be concluded that the majority of empirical studies investigating impact of assessment have mainly focused on large-scale proficiency tests. Despite the fact that teacher assessment practice in a classroom is a complex phenomenon, there are not many empirical studies investigating assessment in a classroom context, compared to considerable literature on large-scale proficiency tests. This concern is expressed by McNamara (2001): 'too much language testing research is about high-stakes proficiency tests, ignoring classroom contexts, and focusing on the use of technically sophisticated quantitative methods to improve the quality of tests at the expense of methods more accessible to non-expert' (p. 329). In

a classroom context, on the other hand, teachers do not necessary have adequate expertise to understand and realise the impact of assessment and their practices on their community and students. In addition, as described earlier in this section noting the variability in assessment practices among teachers, teachers need to be aware of the consequences of assessments of the students.

Furthermore, in a classroom context, as opposed to a large-scale proficiency test context, teachers have to develop assessment, including traditional test and performance-based assessment, for students as well as rate the students' performances. Therefore, they need to realise and take their responsibilities in ensuring the positive impact of the assessment and their practices, especially on students' learning. Since teachers may not have necessary knowledge in assessment to do so, expertise from an assessment professional is needed. It is believed that an on-going in-service professional development programme may be a more productive option to provide teachers with both theoretical and practical understandings of assessment; thus improving teacher assessment practices, which is the main argument of the present study.

3.3 Teacher Change and Professional Development

It has been documented that, historically, teacher change has been directly linked with planned professional development (PD) activity (Clarke & Hollingsworth, 2002). Richardson (1996) reports that before the 1980s, the research into PD focused on teacher behaviours and skills. Later, it began to focus on 'teacher thought processes' with the emphasis on 'the formation or transformation of teacher thinking and reflective processes, dispositions, knowledge, and beliefs' in mid-1980s (p. 110). Richardson also points out that this trend led to a large amount of research which studied the changes of beliefs of teachers at the pre-service and in-service levels. In a

similar vein, Burns (1992) proposes that in designing staff development programmes in a context of language education, it is crucial to 'understand how teachers' beliefs and practices evolve naturally over time' (p. 81). Furthermore, Richardson (op. cit.) reviews the research on changes in teachers' beliefs in staff development programmes and identifies that one of the research areas in teacher change is those studies that examine changes in belief as an outcome of staff development programmes. Richardson reports that 'prior schooling and classroom experiences influence greatly teachers' developing beliefs and knowledge' and 'facilitating meaningful change in both beliefs and practice in in-service teachers may be easier than promoting changes in belief at the pre-service level' (p. 113).

In designing a change study of a collaborative staff development process, its research design should have the following characteristics: open-ended, rich data, multi-method approaches to assessing teacher cognition, presentation to participants of data collected during the staff development process, constructs of change that emerge from the process and data, case studies of individuals and groups of teachers, and a collaborative process (Richardson and Anders, 1994, pp. 165 - 166). In addition, Richardson and Anders (ibid.) argue that reflection and changes are continuous processes of assessing beliefs, goals, and results, and they are thus not static. Therefore, the crucial component is the 'development of a change and reflection orientation to allow the teacher to continue to question both new and old practices' (p. 163). One of the desired results is an awareness of each individual teacher's ways of thinking and instructional practices. In addition to changes in behaviours and actions, the rationale and justifications that accompany new practices are the focal interest of this method (ibid.). Moreover, Kagan (1992, p. 66) concludes about the complexity of teachers' beliefs that:

> Teachers' beliefs appear to be relatively stable and resistant to change and a teacher's beliefs tend to be associated with a congruent style of teaching that

> is often evident across different classes and grade levels. Belief cannot be inferred directly from teacher behavior, because teachers can follow similar practices for very different reasons. Moreover, much of what teachers know or believe about their craft is tacit …

As far as a PD programme is concerned, Guskey (2000) stresses that PD includes the processes and activities which are designed to enhance the professional knowledge, skills, and attitudes of educators so that they might, in turn, improve the learning of students (p. 16). Furthermore, Guskey (ibid., pp. 17 - 22) proposes the following characteristics of PD:

- PD is an *intentional process* designed to bring about positive change and improvement, and guided by a clear vision of purposes and planned goals;
- PD is an *ongoing process*, a job-embedded process, in which educators at all levels must continuously learn throughout the entire span of their professional careers; and
- PD is a *systemic process* that considers change over an extended period of time and takes into account all levels of the organization.

Moreover, PD could be implemented in many forms; action research is one of them (see also Section 5.1.1.5). According to Burns (2005a), action research consists of two components: the action and research. The participants of an action research are involved in ‘a process of planned intervention where concrete strategies, process or activities are developed within the research context’ (Burns, ibid., p. 58) in the *action* component, while the *research* component involves the iterative process of collection of data, data analysis, and reflection on the implications of the findings for further observation and action (p. 59). Thus, action research is, and should be, a highly reflective process. But action research is also a systematic process of investigating practical issues or concerns which arise within a particular social context involving the collaboration of the participants in that context in order to provide evidence that can point to the enhancement of practice, the development of new theoretical

understandings, and the introduction of change into the social context (Burns, 1999, 2005b) – it is highly practical.

Furthermore, action research is contextual, small-scale and localised, the main purpose is to bring about change and improvement in practices, it is a collaborative investigation by a team of teachers and a researcher, and the changes in practice are originated from the data provided by the teachers (Burns 1999, p. 30). In terms of collaboration between teachers and researchers, or the collaborative teacher-researcher project action research (Burns, 2009), Johnston (2009) points out that collaborative teacher development has become an important form of teacher development. He stresses two important features of the collaborative teacher development: teachers concerned must have, or share, control over the process, and the goal of teacher professional development must be clearly stated as the central component of the collaboration (p. 242). However, the significant fundamental challenge pointed out by Johnston in collaborative teacher development is the inequities of power and status. He states that this collaboration internally exhibits an inherent power imbalance in the collaborative relationships between teachers and researchers, for example, a lack of true respect of the researchers on the teachers' contributions.

In terms of investigating PD and teacher learning, Borko (2004) presents a very convincing approach. She stresses that to understand teacher learning from PD courses or workshops, it is important to study it from a situative perspective. This perspective includes a multiple contexts perspective as well as taking into consideration both the 'individual teacher-learners and the social systems in which they are participants' (p. 4). From reviewing numerous studies relating to PD, Borko identifies four elements within a PD system: the PD programme; the teachers, who are the learners in the system; the facilitator, who guides teachers as they construct new knowledge and practices; and the context in which the PD occurs (p. 4). She also

categorises the research on teacher professional development into three phases: phase 1, research activities focus on an individual PD at a single site; phase 2, researchers study a single PD enacted by more than one facilitator at more than one site; phase 3, the researchers compare multiple PDs. However, Borko points out that most studies in PD to date have been in phase 1. The present study is also in this phase. Thus, only phase 1 is presented in this discussion. The purpose of the phase 1 PD activities, a study of a single PD programme at a single site, is to create evidence that a PD programme can create a positive impact on teacher learning. In this phase of PD programme, Borko states that the research provides evidence that 'high-quality PD programs can help teachers deepen their knowledge and transform their teaching'. She also notes that in this PD programme, the designers are usually the researchers themselves and the participants are typically motivated volunteers.

Moreover, Borko discusses what researchers can learn from both the individual and group as the unit of analysis. From individual focus analysis, the findings can reveal how teacher knowledge and practices can change through intensive PD programmes. The research also indicates that meaningful learning is an uncertain and slow process for teachers, some teachers change more than others and, and some elements of teachers' knowledge and practice are more easily changed than others. From group focus point analysis, the findings can reveal how a strong professional community can foster teacher learning. When focused on both individual and group as the unit of analysis, Borko discovered that records of practices are powerful tools for facilitating teacher change. Nonetheless, she stresses that the insights from focusing on either the individual or the group as the unit of analysis are limited in scope. She recommends that researchers, based on a situative perspective, have to use the multiple conceptual frameworks and units of analysis, and have to coordinate them in a manner that leads to a deeper, fuller explanation of teacher development (p. 8).

In summary, in the present study, the activities, carried out as part of the PD in language assessment, are intended to provide teachers with theoretical and practical fundamental understandings of language assessment necessary for them, in their context, to enhance their understandings and improve their assessment practices. Moreover, in designing the study of the impact of the PD, multi-methods in data collection are employed to investigate the development process of the teachers in terms of their views and behaviours in assessment. In the analysis of the data, the emphasis is on both individual teacher and institutional levels.

3.4 Professional Development in Language Assessment

A further significant consideration in language education is the role of availability of teacher professional development. Crandall (2000, p. 36) points out that one of the major shifts in current language teacher education is:

> a growing concern that teaching be viewed as a profession (similar to medicine or law) with respect for the role of teachers in developing theory and directing their own professional development through collaborative observation, teacher research and inquiry, and sustained inservice programs, rather than the typical short-term workshop or training program.

However, in language assessment, teachers in general have always seen testing and assessment as their enemies, or something to be taken care of by the testing experts (Hamp-Lyons, 2003); in consequence, Malone (2008) points out that there is a gap between the training of language teachers in language assessment and language testing practice. The main persisting problem is that:

> there is no consensus on what is required or even needed for language instructors to reliably, and validly develop, select, administer and interpret tests. Therefore, the question remains as to what can be done to support and train those who "have to do the real work of language teaching" (Carroll, 1991, p. 26) when they assess their students. (pp. 225 - 226)

Hamp-Lyons (op. cit.) also emphasises that teachers have to get involved with assessment to a certain extent and have to have enough knowledge about assessment practices to be able to evaluate the assessment being brought into the programs, or being taken externally by the students. She concludes that teachers need to have a 'firm understanding of how assessment works, what it can do, and what it *cannot* do' (p. 183, emphasis in original).

In developing PD programmes in assessment, Brindley (2001) emphasises that it is crucial to know about 'teachers' assessment practices and levels of knowledge' within that particular context (p. 129). Stiggins and Conklin (1993), for example, note that it is important to make certain to correctly match what teachers need to know about assessment and what they are taught about assessment during a training programme, since the inadequate and mismatching of the training has remarkably adverse effects on teachers and the education community in general. In understanding teachers' professional development, Tsui (2007) also comments that 'the interaction between teachers' conceptions of teaching and learning and their world of practice is an important dimension that should be taken into consideration' (p. 1055).

Brindley (op. cit.) recommends that a PD programme in language assessment should include the components of social context of assessment (core unit); defining and describing proficiency (core unit), constructing and evaluating language test, assessment in the language curriculum, and putting assessment into practice. Moreover, he suggests that it should involve the whole system, capitalise on existing practices, recognise and deal with the reality and constraints influencing teachers'

assessment practices, encourage a research orientation to PD, and plan for change. Brindley also points out that the implementation of the PD programme could be done in a modular fashion, in the form of a short course, series of seminars/workshops or individual seminar/workshops. Similarly, Malone (2008) proposes that the first step is to determine what teachers need to know about language assessment in order to perform their jobs, and secondly, to determine how to provide such training. She also stresses that it is very crucial to identify gaps: 'what do instructors know about assessment, what do they need to know and how can this information best and most effectively be shared?'(p. 237). In addition, in order to gain a greater insight into the actual state of professional knowledge and practices surrounding language testing, studies using more observational, ethnographic or longitudinal data are encouraged.

However, it should be noted that training in a workshop format can be time-consuming, expensive, and limited in its ability to reach all language teachers (Malone, ibid.). Nevertheless, Malone agrees that language teachers should participate in a regular in-service training to supplement the pre-service teacher training program they have had because an ongoing PD can 'keep teachers abreast of current developments in language assessment and allow them to apply new development to the language classroom (p. 236). In the same vein, Hamp-Lyons (2002) recognises the development of 'the fourth generation' of assessment of writing which 'will need to be technological, humanistic, political, and ethical'. With the development of technology, she points out that it is crucial to 'empower not only large test agencies, but more importantly test-takers, raters and educators' (pp. 12 - 13). Hamp-Lyons (2007b, p. 499) proposes that:

> A far more fruitful way into professional development for teachers is to involve them in performance assessment judgments and rater training (Hamp-Lyons & Condon, 2000). Since teachers are both interlocutors and raters in their own classrooms, professional development can capitalize on the variability of response to language performances and help teachers to, first, deconstruct their own preferred ways of responding to learners' language, and then to establish a consistent approach to responding to student work.

When teacher-assessors receive adequate preparation, Hamp-Lyons also emphasises that apart from being more self-consistent in the assessments in their classroom, the teachers will have opportunities to 'critique their position in the education society, identify points of opportunity and mechanisms to influence education planning, including assessment, and to find ways to contribute to positive change' (p. 492). When teachers get together, arguments, understandings, clarifications, and interpretations are constructed though discussion with other teachers (Mann, 2005, p. 111). Malone (2008), likewise, agrees that the major goal of training in language assessment is to empower language teachers. In addition, the training will 'improve the language assessment being conducted and promote positive washback to teaching and learning' (p. 237). With available resources, especially textbooks, implementing training in language testing and assessment should be more practical for language educators (for detail of textbooks in teaching language testing, see Davies, 2008).

In the present study, a series of in-service PD in language assessment was carried out for Thai EFL teachers who did not have a background in language assessment, but were responsible for assessment in the classroom and institutional levels. One of the major objectives of the PD was to create a positive impact on the teachers who participated in the programme.

3.5 Innovation Theory and Teacher Change

It has been asserted that innovation and change have become a necessary part of teacher development (Mann, 2005). Wall (2000), from having applied educational

innovation theory in examining the impact of high-stakes examinations on classroom teaching (as described above), observes that educational innovation frameworks yield valuable insights in investigating studies in language testing and assessment, in particularly, changes. The table below summarises her observation. In the sub-sections below, I present different views of implementation/ diffusion of innovation because different views have different implications in investigating the impact of the PD in the present study.

Table 3.1: Characteristics of innovation from language assessment perspective (adapted from Wall, 2000, pp. 503-504)

Related to adoption of innovation
The users of an innovation will reach different 'levels of implementation'.
Every innovation has a number of characteristics, some of which may facilitate its adoption and some of which may hinder it.
It is necessary to analyse the context of an innovation in order to judge whether it is likely to be adopted.
The rate of adoption of an innovation is determined by many factors.
Related to process of innovation
The process of innovation is long and complex, consisting of many stages
There are many participants involved in the process of innovation, each with their own needs and limitations.
The meaning of an innovation will be different for every individual involved in the process.
Related to change
Innovation is different from other sorts of change.
An innovation may require change on three different levels: content, methodology and attitudes.
It is difficult to measure some kinds of changes, especially changes in awareness or changes which are open-ended.
There are a variety of models for introducing change.
It takes time before an innovation can bring about fundamental changes.

3.5.1 Rogers' view

Drawing from the insights from various disciplines, such as agricultural innovations, educational innovations, health and family planning innovations, Rogers (2003, p. 12) defines the term innovation as:

> an idea, practice, or object that is perceived as new by an individual or other unit of adoption ... The perceived newness of the idea for the individual determines his or her reaction to it. If an idea seems new to the individual, it is an innovation.

The process in which 'an innovation is communicated through certain channels over time among the members of a social system,' is called the 'diffusion'. The diffusion of innovation, which is a two-way process of convergence, involves the

communication of a new idea in which participants create and share this new idea with one another in order to reach a mutual understanding. Moreover, Rogers points out that diffusion is a kind of 'social change' by which change happens in the 'structure and function of a social system. When new ideas are invented, diffused, and adopted or rejected, leading to certain consequences, social change occurs' (p. 5). In other words, diffusion 'is the process by which (1) an *innovation* (2) is *communicated* through certain channels (3) *over time* (4) among the members of a *social system*' (p. 11, emphasis in original).

Furthermore, Rogers proposes six main stages in the innovation-development process: recognising a problem or need, basic and applied research, development commercialisation, diffusion and adoption, and consequences. Rogers, however, points out that 'the six stages may not always occur in a linear sequences, the time order of the stages may be different, and certain stages may not occur at all' (p. 167). Once an innovation has been developed, it depends on an individual or a system to make decision whether or not to incorporate the innovation into ongoing practice that is the 'innovation-decision process.'

Rogers (p. 169) proposes five stages of the innovation-decision process:

1 *knowledge*, which occurs when an individual (or other decision-making unit) is exposed to an innovation's existence and gains an understanding of how it functions;
2 *persuasion*, which occurs when an individual (or other decision-making unit) forms a favourable or an unfavourable attitude towards the innovation;
3 *decision*, which takes place when an individual (or other decision-making unit) engages in activities that lead to a choice to adopt or reject the innovation;
4 *implementation*, which occurs when an individual (or other decision-making unit) puts a new idea into use; and

5 *confirmation*, which takes place when an individual seeks reinforcement of an innovation-decision already made, but he or she may reverse this previous decision if exposed to conflicting messages about the innovation.

3.5.2 Fullan's view

In the education context, Fullan (2004, p. 65), in his 'innovation-focused' approach, identifies three broad phases of the change process of which the outcomes pose the question of whether or not student learning is enhanced, and whether or not experiences with change increase subsequent capacity to deal with future change:

Phase I – initiation (or mobilization, adoption) – consists of the process that leads up to and includes a decision to adopt or proceed with a change.

Phrase II – implementation (or initial use) (usually the first 2 or 3 years of use) – involves the first experiences of attempting to put an idea or reform into practice.

Phrase III – institutionalisation (or continuation, incorporation, routinisation) – refers to whether the change gets built in as an ongoing part of the system or disappears by way of a decision to discard or through attrition.

Fullan emphasises that 'what happens at one stage of the change process strongly affects subsequent stage, but new determinants also appear', and 'the three phases should be considered at the outset' that is 'the moment that initiating begins is the moment that the stage is being set for implementation and continuation' (p. 69).

Fullan also identifies eight sources affecting the initiation stage: existence of quality of innovations, access to innovation, advocacy from central administration, teacher advocacy, external change agents, community pressure/support/apathy, new policy – funds (federal/state/local), and problem-solving and bureaucratic orientations (p. 70). Moreover, he lists nine critical factors affecting the implementation phase, which can be grouped into three main categories:

1 Characteristics of change; including need, clarity, complexity, and quality/practicality

2 Local characteristics; including district, community, principal, and teacher

3 External factors; including government and other agencies (p. 87)

3.5.3 Markee's view

In the diffusion of innovation in language education, Markee (1997) provides principles for language teaching professionals to understand the factors that affect the design, implementation, and maintenance of innovations. His framework is based on the questions posted by Cooper (1982; 1989): 'who adopts what, where, when, why and how?' (p. 118). In terms of 'who', Markee, based on Fullan (1982), points out that 'teachers are key players in all language teaching innovation; however, many other individuals also have a stake in the innovation process'. Though the participants in the innovation decision process are different from context to context, they tend to 'assume social roles that define their relationships with other stakeholders (p. 43). Markee also reports Kennedy's (1988) study that these individuals consist of Ministry of Education Officials, Deans, or Heads of Department who take the role of *adopter*; teachers are *implementers*; students are *clients*; curriculum and materials designers are *suppliers*; and the expatriate curriculum experts are the change agent.

In the decision-making processes of potential adopters, drawing from the studies by Rogers (1983) and Rogers and Shoemaker (1971), Markee identifies four phases:

1 Gaining knowledge about an innovation

2 Being persuaded of its value

3 Making a preliminary decision whether to adopt or reject the innovation and implementing this decision

4 Confirming or disconfirming their previous decision (p. 45)

In terms of 'what', Markee defines curricular innovation as 'a managed process of development whose principal products are teaching (and/or testing) materials, methodological skills, and pedagogical values that are perceived as new by

potential adopters' (p. 46). Under 'where', Markee cites Cooper (1989) that '*where* in an innovation is implemented is a sociocultural, not a geographical, issue' (p. 55, emphasis in original). Drawing from Kennedy (1988), Markee reports that in managing the implementation of curricular innovation, a change agent must take into consideration the following factors: classroom innovation, institutional, educational, administrative, political, and cultural. Under 'when' Markee points out that the rate of adaptation varies. Markee argues that 'the diffusion process tends to begin slowly; it then suddenly accelerates and finally slackens off' (p. 58). He also stresses that innovation takes time to implement and always takes longer to implement than expected.

In terms of 'why', Markee points out that the first factor where innovations are adopted is within the sociocultural constraints (as described in the 'what' section above). The second reason is the different psychological profiles of the adopters, for example, early adopters tend to be personally or professionally close to change agents and are often willing to take risks. Finally, another factor affecting the adaptation of innovations is the attributes of the innovations themselves. Drawing from Rogers (1983), Markee reports five attributes to the decision to adopt or reject an innovation:

1 The *relative advantages* of adopting an innovation – the costs or benefits
2 Its *compatibility* with previous practice – how different or similar the innovation is to what the potential adopter already uses
3 Its *complexity* – how difficult the innovation is to understand or use
4 Its *trialability* – how easy it is to try out in stages
5 Its *observability* – how visible the innovation is

Finally, under 'how', Markee describes five different approaches to affecting change: the social interaction model; center-periphery model; research, development, and diffusion model; problem-solving model; and the linkage model.

3.5.4 Henrichsen's view

From his attempt in diffusion of innovations in English language teaching by the English Language Exploratory Committee in Japan, Henrichsen (1989) proposes a 'Hybrid Model' of the diffusion/implementation process. The model consists of three main elements: antecedents, process and consequences.

The 'antecedents' section of the hybrid model focuses on the significance of investing the historical nature and development of the following four influential factors as part of the planning process. These factors are:

Characteristics of the intended-user system, including the structure and power relationships in schools and society;

Characteristics of the intended users of the innovation, including their attitudes, values, norms, and abilities;

Traditional pedagogical practices, deriving from different cultural and historical practices in teaching and learning; and

Experiences of previous reformers, providing a knowledge and understanding of how to achieve the goals or how prepare for potential difficulties.

In the 'process' component, Henrichsen points out the significant roles of analyzing the factors which influence the change process. He provides the factors that may hinder or facilitate change within each element:

- *Within the innovation itself*, including originality, complexity, explicitness, relative advantage, trialability, observability, status, practicality, flexibility/adaptability, primacy, and form;
- *Within the resource system*, including capacity, structure, openness, and harmony;
- *Within the intended-user system*, including geographic location, centralisation of power and administration, size of the adopting unit, communication structure, group orientation and tolerance of deviancy, openness, teacher factors, learner factors, capacities, educational philosophy, and examination ;and

- *Inter-elemental*, including compatibility, linkage, reward, proximity, and synergism

For 'consequences', the hybrid model provides different types of the innovation decisions and the outcomes. There are three types of innovation decisions: *optional* decisions – an individual may choose to implement an innovation independent of the decisions made by other members of society; *collective* decisions – the decisions are made only by consensus agreement among all the parties involved; and *authority* decisions – the decisions are forced upon individuals by someone in a superordinate power position; and *contingent* decisions – the decisions are chained to others, made only after a prior decision and depend on the nature of that decision. The results of implementing innovation may be immediate or delayed; be direct or indirect; be manifest or latent; functional or dysfunctional or both; and have desirable or undesirable effects.

3.5.5 Summary

Based on the discussions above, different theory provides different implications for the present study in examining the impact of the implementation of innovation. Rogers (2003) provides general definitions of innovation, its basic characteristics, and its diffusion process. Fullan (2007), on the other hand, proposes broad phases of change process in educational innovation as well as factors affecting the each phase. A more specific view of educational innovation from a language educator perspective is provided by Markee (1997). Markee's framework on who adopts what, where, when, why and how proved to be very helpful. Finally, Henrichsen's (1989) hybrid model of the diffusion/implementation process offers insights into factors affecting different stages of innovation.

3.6 Conclusion

In this chapter, I have discussed the concept of teacher change with the emphasis on the studies of teacher change in language testing and assessment. I also investigated the concepts in teacher change in relations to PD, as well as some theories from the field of educational innovation. From this investigation, I have found that teacher change is a long and complex process. Also, change can be in many forms, such as change in attitude, awareness, and behaviour. However, teachers may or may not change at all after having participated in a PD designed to create positive change. Therefore, the present study was designed as a longitudinal study to understand the development of teachers who participated in the PD in language assessment. As indicated in the literature, before the implementation of a PD, a thorough context investigation is needed. Thus, I carried out a pilot study to gain an in-depth understanding of the research context before the main study. The findings from the pilot study are presented in the next chapter (Section 4.2). The data from the main study and the follow-up study will be explained in Chapter 6. The discussion of the findings will be explored in Chapter 7.

4 Understanding the Context under Investigation

The present study is embedded in the context of teaching and testing in Thailand. It is, therefore, important to understand the country's history and present situations in teaching and testing as they, directly and indirectly, facilitate in understanding of the participants in the study. In this chapter, I firstly explore the historical aspect of English teaching and testing in Thailand. The second part of this chapter reports the findings from the pilot study, which was conducted at Chiang Mai University. The main purpose of this study was to gain an in-depth understanding of the context focusing on the beliefs, attitudes, knowledge, and practices related to assessment of the teachers in the English Department. The first part of the discussion is the findings from the observations and the second part from the case study of five teachers.

4.1 English Language Teaching and Testing in Thailand

Thailand uses English primarily as a lingua franca or foreign language for international relations and business. English is the de facto second language, although there is no *official* second language, and the language is used in a wide range of domains. Moreover, English has been recognized as a crucial skill for 'professional advancement' in urban areas (Baker, 2008). Foley (2005) agrees that in Thailand, 'English proficiency offers opportunities and access to technology, communication and professional advancement' (p. 227).

Wongsathorn, Hiranburana and Chinnawongs (2002) trace English language teaching (ELT) in Thailand and report that ELT in the country dates back to the reign of King Rama III (1824 - 1851 A.D), but English was only available to a small group of people. English did not become a compulsory subject for students beyond grade

four until 1921. In 1960 there was a change in the English syllabus for secondary schools, in which the four language skills were given equal emphasis. The aim of English teaching was to enable students to use the language for international communication and for acquiring knowledge and information. Wongsathorn et al. adds that there were attempts to replace rote memorisation and grammar translation, the traditional methods, with the audio-lingual teaching methods but without much success.

In the 1977 and 1980 curricula, foreign languages were classified as electives to be taught in secondary school nationwide. At the tertiary level, six credits of language were required as part of general education, with English being the most popular required foreign language. The curricula aimed to enable students to use English primarily for communicative purposes in all four skills. However, Wongsathorn et al. (ibid.) point out that there was a lack of qualified teachers in most primary schools. Foley (2005), nonetheless, reports that this was the period when the British Council were involved in running a series of in-service courses for teachers to help with this problem. In 1996, English was made compulsory for all primary students from grade one onwards. The emphasis of this revised curriculum was the development of students' language proficiency for the purpose of communication, acquisition of knowledge, use of English for academic purposes, career advancement, and appreciation of the language and its culture. The teaching approach could be described as 'functional-communicative with an eclectic orientation' with learner autonomy at its central focus (Wongsathorn et al., op. cit.). The assessment consisted of portfolios, records and observation, and formal assessment.

With the 1999 National Educational Act and the Ministry of University Affair's announcement of a new policy on English Instruction of Liberal Education in the year 2000, English, together with IT skills, has been placed at the 'forefront of national intellectual development' (Wongsathorn et al., ibid.). According to this Act,

the revised English curricula in Thai tertiary institutions required at least twelve credits instead of six credits as required earlier in each of the following: six credits in general English and six for English for academic or specific purposes. The emphases are on autonomous learning, innovations and new technology, and performance standards.

As far as assessment is concerned, though the English syllabi have changed, for example, the 1996 syllabus focuses on the functional-communicative language (Wongsathorn et al., ibid.), Prapphal (2008) states that language testing has not changed. Multiple-choice is still the most common test format and the majority of the tests still target the functions and structure of the language. One of the reasons reported is that teachers do not have the time to grade essays or implement continuous assessment (Foley, op. cit.). Although, there was an educational reform in 1999, language testing practice was not part of the change (Prapphal, op. cit.). Moreover, Prapphal (ibid.) reports that washback effects of language tests (for more detail on washback, see Section 3.2.1) have been one of the main assessment issues in Thailand. She observes that in many schools the teaching and learning process in the last semester of the last academic year, before the university entrance exams, focuses on 'reviewing the content and format' of the exams (p. 129). The same problem has also been recognised by other scholars. For instance, Wongsothorn et al. (op. cit.) point out that the high stakes university entrance exams which only examine reading skills and grammar knowledge have led to a neglect of productive skills in the classroom. These skills, moreover, have never been included in testing in high-stakes exams.

Nevertheless, there have been attempts in changing assessment practices in tertiary education, for example, the implementation of a task-based assessment approach (McDonough & Chaikitmongkol, 2007; Watson Todd, 2006). Watson Todd (ibid.) investigated the changes of the task-based curriculum during its four years of

implementation at King Mongkut's University of Technology in Thonburi. From the interviews with the teachers and the course documentation, Watson Todd found that the courses assessment changed its focus from continuous assessment to the increased emphasis on examinations because of the rater reliability issues. Watson Todd reports that there were teachers, largely part-time teachers, who did not follow the set criteria when marking assignments. He adds that these teachers were not fully inducted in to the introduction of the task-based assessment. To solve these reliability problems, the course team decided to increase the proportion of marks given to exams. For the McDonough and Chaikitmongkol's (2007) study, see Section 4.2 below.

As far as teacher education is concerned, in line with the National Education Act of 1999, the Ministry of Education of Thailand aims to promote pre-service as well as in-service development schemes and activities in all levels of education. For example, many organisations with sufficient funds supporting staff development were set up (Commission on Higher Education, n.d.). However, according to the report prepared for the Office of the National Education Commission and the Asian Development Bank, there is a gap between 'the level of knowledge and practices of Thai educators and their institution on one hand, and the necessary level of knowledge, skills and practices of people' on the other hand (Pillay, 2002). This gap results from:

> a lack of cross-institution dialogue and investment in education, those responsible for teacher training and development – Thai educators and their institutions – have not made enough effort to provide the necessary leadership in understanding and implementing educational reform and teacher development. Further, they have not routinely participated in international 'learning communities' or been involved in or become familiar with innovative research in teacher development. (Pillay, ibid. p. 8)

Pillay (ibid.) concludes that despite the Government's efforts in providing necessary physical resources and infrastructure to provide pre-service and in-service, the quality

of teacher training and development is increasingly becoming a concern for all stakeholders in Thailand because 'the quality of teachers and education in general continues to decline' (p. 9).

Furthermore, in language testing and assessment, teacher education in this area is very limited because training programmes in language testing 'are accessible to only a small proportion of language teachers' (Prapphal, 2008, p. 136). With the increased complexity in the assessment systems required by the educational reform (with the emphasis on self-assessment and peer-assessment), Prapphal (ibid.) stresses that in order for the reform to be successful, Thai teachers need improvement in their language assessment knowledge. For instance, language teachers should be able to ensure the reliability, validity and practicality of their assessments. She proposes that 'the long-term success of the National Education Act may well depend on the ability of teachers to change the way they conduct language testing' (p. 136).

4.2 English Language Teaching and Testing at Chiang Mai University: The Pilot study

The main purpose of this pilot study was to gain an in-depth understanding of the context focusing on the beliefs, attitudes, knowledge, and practices related to assessment of the teachers in the English department. The understandings of the context in the pilot study would indicate the directions of the main study. In addition, different research methods were used in order to find out the most suitable methods for the main study.

This study was carried out at the English department, Faculty of Humanities, Chiang Mai University. Chiang Mai University, the first provincial university in Thailand established in 1964, is a large public university in the north of Thailand. The English department is one of the largest departments in the University because it has to teach English major students (and about 100 new students every year) and is

additionally responsible for the FE courses required for every student (approximately 20,000 students a year). There are approximately 70 full-time and more than 30 part-time teachers. In 2002, the University, responding to the new policy on English Instruction of Liberal Education (see Section 4.1), requested the English department to revise the existing foundation courses, which were a focus-on-form approach. After a questionnaire-based needs analysis, it was found that teachers and students were not satisfied with the previous English courses. Therefore, the Department reviewed relevant literature in order to plan for the new courses and proposed six courses (Winitchaikul, Wiriyachitra & Chaikitmongkol, 2002). These courses were Foundation English (FE) 1 (of which I was one of the three material/assessment writing team), FE 2, English for Academic Purposes, and English for Specific Purposes (including 3 different courses: Social Science and Humanities, Science and Technology and Health Science). The focus of the present study is FE 2.

FE 1 and FE 2 followed an integrated-skills task syllabus with the incorporation of learning strategies into the course. There are 3 tasks for each course (for an example task, see Appendix E). Each task, comprising of writing and speaking components (in this thesis the term 'task' refers to this writing or speaking task) requires approximately eight 75-minute class periods to complete. The courses incorporate listening and reading materials from a commercial textbook, *Skyline 3* (Brewster, Davies, & Roger, 2001). The excerpts from the textbook were selected to complement the tasks' content as well as the knowledge and skills needed to carry out the tasks. The courses also includes supplementary materials (*Student's Workbook*) created by the course material writing team. The teachers were provided with a *Teacher's Guide* describing in great detail how each class period should be spent. The *Teacher's Guide* included the task objectives, class objectives, and objectives of the class activities. Moreover, the *Teacher's Guide* provided the teachers with suggested teaching procedures and notes/ tips/ suggested answers to the questions in the activities. For the assessment of the course, see Section 4.2.3.1 (see also Table 4.1).

McDonough and Chaikitmongkol (2007) investigated the reactions of 13 teachers and 35 students toward FE 1 and found that both teachers and students had positive reactions to the course. From the rich qualitative data, including task evaluation, learning notebook, classroom observation, course evaluation, teachers and students' interviews, and field notes, the findings revealed that both teachers and students believed that the course encouraged autonomous learning and 'real world academic needs,' though there were initial negative reactions toward the lack of explicit grammar instruction in the course content. Though the tasks, the focus of the course, were themselves the assessment, McDonough and Chaikitmongkol (ibid.) did not explore the reactions of teachers or students toward the assessment aspects of the course in their study. Therefore, the study on the tasks' assessment was carried out in the pilot study to shed some lights on the assessment dimension of the course.

4.2.1 Research design

Being a pilot study, one of the purposes of this study was to try out different research methods and decide on the appropriate ones to be adopted for the main study. The main purpose of the study was to gain thorough understanding of the context under investigation as a preparation for the main study.

4.2.1.1 Purposes of the study

In order to acquire a comprehensive understanding of the research context, especially the needs and problems concerning assessment, it was important to:

1 Understand teachers' beliefs, attitudes and knowledge in language assessment

2 Understand the relationships between the above constructs and what teachers do

3 Find out the needs and problems related to assessment in the Department

4.2.1.2 Research questions

This study was guided by the following research questions.

1 What kinds of assessment are being used at the Department and what are their characteristics?

2 What are the beliefs, knowledge and attitudes of Thai teachers at the Department towards the assessment being used and why do they have those beliefs and attitudes?

3 How do teachers perform assessment in their classroom and how do their beliefs, knowledge and attitudes affect their practice in assessment?

4 What areas of assessment require attention in order to improve teacher's practice in assessment?

5 How a professional development programme could be implemented?

4.2.1.3 Data collection processes

The process of collecting the data was divided into three stages. Each stage had specific purposes and different research methods were used for each purpose.

First stage

The initial objective of the first stage was to obtain sufficient background information in order to understand the context and identify the problems the Department has had with testing and assessment for the foundation courses. The followings methods were used to collect the data:

- Review of the documents related to assessment such as the history of assessment used, past seminars/ workshops/ training in assessment, and complaints from the students.
- Semi-structured interviews used to obtain more understanding of the context and problems. The participants were those directly involved in assessment as well as administrators, including test developers, test item writers, the chief coordinator

of the foundation courses, and the coordinator of FE 2. The interviews were done in Thai and audio recorded. They were later transcribed and translated.

Second stage

The objective of the second step of the study was to explore the assessment beliefs, knowledge, and attitudes of the teachers at the Department. Another aim of this stage was to identify the needs and problems of teacher assessment. In order to acquire the overall information, questionnaire surveys were used and semi-structured interviews were later used for more in-depth information.

- The participants of the questionnaire survey were all teachers at the Department who teach FE 2. The questionnaire included questions eliciting the bio-data of the teachers, their assessment beliefs, attitudes, knowledge, and needs in assessment. It should be noted that after the field work, conducting analysis of the data and further extensive literature review, I have decided to adopt qualitative research methodology for the present study (for more detail, see Section 5.1). Thus, the analysis of the questionnaire survey is not reported in this thesis and a questionnaire survey would not be employed in the main study.
- The semi-structured interview included 5 participants who currently taught FE 2. The main purpose of the interviews was to obtain more insight into their assessment beliefs, knowledge, and attitudes as well as their needs in assessment. The interviews focused on the teachers' views of the assessment tasks, assessment process, assessment products, and assessor needs. The interviews were done in Thai and audio recorded. They were later transcribed and translated. As grounded theory was employed as the tool for data analysis, the questions used in the interviews aimed at encouraging the interviewers to unfold their beliefs, attitudes and experiences.

Third stage

The final phase of the pilot study analysed the teachers' assessment practices by focusing on how they do assessment and why they assess in that particular way.

- The first method used was classroom observation. The same group of teachers who had been chosen for the in-depth interviews were also participants in the classroom observations. Activities in the classroom were audio recorded apart from the observation field notes.
- After the observations, follow-up interviews with the teachers being observed were conducted. The major aim of this introspective interview was to invite the teachers to share their views of the way they do certain things in the classroom relating to assessment. Moreover, stimulated verbal method was employed in the interviews. The teachers were provided with the assessment tasks they had marked or rated which included copies of written tasks, audio clips of oral tasks, and final exam papers of the students. Three tasks of each type of assessment were used: one from each performance level (high, average and low scores). Teachers were asked to comment on their thought processes along the way. Furthermore, at this stage, I had already done some analysis of the previous interviews, and was able to ask teachers for clarification on points which were unclear from previous interviews.

4.2.1.4 Participant profiles

The participants in the pilot study included 5 teachers who were teaching FE courses at the time of the study. Though they were selected with opportunity and convenience taken into account, I was successful in recruiting participants from different backgrounds, such as gender, education, and teaching experience. It should be noted that the participants in the pilot study are not the same teachers in the main study (for the participants in the main study, see Section 4.2.1.4).

Teacher 1: Arkom is the youngest of the participants (30 years old) only having worked at the Department for just over 3 years. However, he has been teaching EFL for a few years before joining the Department. Arkom holds an MA in English and Communication from Chiang Mai University, Thailand. He is the assistant coordinator of FE 1. He was also the assistant when I was the coordinator of this course. I worked with him for one semester before taking my study leave.

Teacher 2: Wawan is 38 years old and the least experienced teacher participant only having taught EFL for 3 years. This is her first teaching job after graduating with a Masters' degree in Education (TEFL) from Chiang Mai University, Thailand. Similarly to Arkom, Wawan was an assistant coordinator of FE 1 for one semester while I was the course coordinator (before she took a maternal leave and Arkom replaced her).

Teacher 3: Muun is 56 years old, the oldest and most experienced teacher who participated in this study. She has a Masters' of Education in TEFL from Chiang Mai University, Thailand. She is in her last year of a PhD in Curriculum & Instruction (full-time), also from Chiang Mai University. She has been teaching at the Department for 30 years.

Teacher 4: Ronnie is 35 years old and has been teaching at the Department for 10 years. He holds an MA in English Literature from Chulalongkorn University, Thailand. Ronnie is an assistant coordinator of FE 2 (the focus of this study). He was also part of the team who wrote the materials for the course. In addition, he has been a member of the exam committee.

Teacher 5: Pawida is 54 years old and has been teaching at the Department for 29 years. She is an assistant professor with a Masters' degree in Science (Curriculum & Instruction) from Baylor University, USA. She has also attended a training programme in language testing and assessment at the University of Cambridge Local Examinations Syndicate, UK. Pawida was the one of the FE 1 material writers. In

addition, she has been the consultant of all four new foundation courses. She was also one of the teachers who taught FE 1 and 2 during the pilot periods.

In the following sections, I will report the findings from this study. The findings are divided into two parts: the findings from the observations and from the case study of five teachers. It should be noted that though different research methods were used to in this study, the findings of the methods relevant to the conclusion and implications for the main study are reported; including the findings from field observations and interviews.

4.2.2 Findings from observations: Assessment practice in the Department

In this section, I report the assessment practices in the Department from my observations. The findings include the FE 2 assessment, standardisation and the Department grade meeting.

4.2.2.1 Assessment practice in Foundation English 2

The table below shows the course assessment of the second semester 2006. However, when the course was first implemented, the course evaluation was different. It included student attendance (10%), performance-based assessment (24%), in-class work (16%) and a traditional test (50%). In addition, the Department also changed the testing approach from a norm-referenced to a criterion-referenced approach in interpreting the students' scores. However, the feedback was reported using the traditional grade system (A, B+, B, C+, C, D+, D, and F). In 2005, however, after being used for the first time over one year, the course evaluation was changed. The grade percentage of a traditional test was increased to 56%, performance-based assessment stayed the same (24%), students' attendance remained constant at 10%, while in-class work was replaced by self-access learning (10%). For class attendance, weighted at 10% of the course assessment, two percent was deducted per class period

missed. The second part of the assessment was self-access learning. Students who took the first two FE courses were required to do self-access learning by accessing E-SALL (Electronic Self-Access Language Learning) via the Internet both on and off-campus. The E-SALL included exercises relating to the lessons and tasks. Students could gain up to 5% from doing the exercises online and another 5% from classroom quizzes based on the online exercises.

Table 4.1: Course assessment

Attendance	10 %
Self-Access Learning	
Online	5%
In-class quizzes	5%
Performance-based assessment (Written and Oral tasks)	
Task 1	8%
Task 2	8%
Task 3	8%
Final exam	56%

Another crucial part of the assessment was the performance-based assessment, which was weighted at 24% of the total. As mentioned above, the course followed a task-based syllabus: the performance-based assessment was an end product of the task culminating with a written assignment and an individual or group oral presentation. There were three tasks (i.e. 3 written reports and 3 oral presentations). Each assessment was weighted at 4%. The most weighted part of the assessment, 56%, was the final exam. The exam was designed to be an achievement test with tasks similar to those taught in class. The item types also resembled classroom activities so that students would be familiar with the format of the exam.

It should be noted that the situation in this institution, of which the weight of final exam was increased after one year of implementation, is very similar to the one reported by Watson Todd (2006) described in Section 4.1 where the course team decided to increase the proportion of marks given to exams to solve the reliability

problems. Rater reliability, resulted from teachers who do not follow the rating criteria when rating performances, seems to be one of the major problems with performance-based assessment in Thailand. The reason teachers do not follow the criteria could be because they are not provided with sufficient and adequate information about the course, its assessment tasks, and the rating criteria. In other words, they do not receive effective preparation or training.

4.2.2.2 Standardisation of the assessment: rater training

Since there are more than a hundred sections and over 70 teachers teaching the foundation course each semester, it was difficult to control subjectivity in grading the tasks. Thus, the Department decided to conduct a standardisation project in order to ensure the inter-rater reliability. However, I would call it a standardisation attempt because of some its misleading procedures (cf. Section 2.4.2). The team distributed two randomly chosen samples of students' written tasks to all teachers. The teachers were asked to assign scores to the sample tasks using the given criteria. Then, the team would calculate the mean score from the scores given by these teachers. The consensus score (or the mean score) would be announced. The teachers were asked not to grade their students' written work before getting this mean score. They were then told that the mean score was the score to keep in mind while grading written tasks with the same quality.

Though the Department tried to make certain of inter-rater reliability, the misled standardisation procedures did not help increase the reliability of the rating. Moreover, the standardisation attempt was done for a written task only. No attempt was carried out at all for the oral tasks. It can be said that there was not any training for teachers to prepare themselves to rate students' performances. Nonetheless, at the beginning of the first semester of 2007, after the pilot study had been conducted and I had already left the research site, the Department provided one day of standardisation training in which teachers had the opportunities to rate samples of students' written

and oral tasks of FE 1 and discuss results with coordinators and colleagues (FE 1 coordinator, personal communication, July 2007).

4.2.2.3 The Department grade meeting

The staff in the grade meeting included the head of the Department, the Department committee and the course coordinators. The course coordinators explained the assessment used in their courses. Then they reported the cut scores based on the previous academic year. Next they reported the numbers of students receiving each grade using those cutting scores. The basic statistical information including mean scores, standard deviations, modes, and medians were given. The committee studied the figures briefly and made some comments if they saw anything unusual.

There were arguments on whether the cut scores for the FE courses should be the ones from the previous academic year. Since there were not substantial reasons for changing the cut scores, the ones from the previous year were used for all the foundation courses. There were also reports on too high scores in some sections in the foundation courses. The high scores in some sections affected the grade cutting procedures. In one case, the coordinators deducted 4.6 points from all students in two sections of one particular teacher because the scores of all students in those two sections were too high.

From the observation from the grade meeting, teachers, who did not have strong background in language testing and assessment, based their assessment practices on what had been laid out for them from the previous years. They did not want to challenge these conventions even they did not agree with. For instance, though they agreed that it was not right to use the same cut scores from the previous year, they could not change the cut scores adopted in the present semester because had no evidence to argue whether the final examination of this academic year were at the similar level of difficulty from the previous year. Nor they were certain of the reliability of teachers in their ratings of performance-based assessment, as

demonstrated by the fact that they had to deduct points from all students in the sections with too high scores. Moreover, some teachers pointed out that it was not fair that the coordinator deducted points from these students since they did have any knowledge about these students.

4.2.3 Findings from case studies: Teachers' views toward the assessment

In this section, I report the views of individual teachers, who participated in the pilot study, toward the assessment of the FE 2, including writing assessment, oral assessment, and final examination, as well as their reported assessment practices.

4.2.3.1 Thinking about the course's assessment in general

Arkom

Arkom stated that he did not like the multiple-choice exam. He said that "*it's useless*" because he did not agree with the idea of treating the performances of students as black and white the way the multiple-choice exam did. He proposed that teachers should not judge students' performances as black and white. Arkom stated very strongly his preference of performance-based assessment. He said, "*I believe in performance-based assessment, for example, a role-play or an oral presentation – anything where students speak*". He added that he preferred oral to written assessment. He believed that that the main objective of the course was spoken performance and not written performance. Therefore, the target of the course was oral performance and thus what students should have mastered after completion of the course. Regarding the rating criteria, Arkom stated that his attitude was one of indifference. Based on his experience as the assistant coordinator of FE 1, he concluded that the criteria could not satisfy every teacher. However, from his point of view, he thought that the course and its materials including the criteria were good enough because they had been piloted and revised thoroughly by the committee.

Concerning the role of teachers as assessors of students' performances, Arkom stated very clearly that teachers were not the judges and should not think that they were.

Wawan

Wawan wanted to assess the students of the FE courses using interviews. She pointed out that most of the time in her classes; students did not speak even with her encouragement. Wawan said she wanted the students to speak and to raise different issues in class and have students discuss them, which was more than what they did at the time in the course. Moreover, Wawan, from her previous education, was aware that there were many ways to assess in a task-based course. However, she understood that because of the constraints of the Department such as increasing workloads with authentic assessment, the assessment was not truly authentic because it did not involve what students had to do in daily life, but at an acceptable level.

Regarding the increase of workloads of the performance-based assessment, Wawan pointed out that many teachers agreed that "[it's] *like we teach a writing course. There are written tasks in these new foundation courses*". She thought that the course emphasised too much on writing. With more assignments to rate, Wawan felt frustrated as it took her a lot of time to score students' written tasks because she gave detailed comments. Moreover, fairness was very important in assessment, Wawan added. She said that in the performance-based assessment she had to be fair with the students when she rated their performances. She stressed that when she did the scoring, "*There's no bias, for instance, I like this student* …" Wawan also pointed out that in order to be fair, when she rated the students' performances, she followed "*the given criteria not on my judgement*". However, she admitted that because of the different levels of student abilities in one class, in some occasions she had to re-consider the scores she had already given to the students. Wawan, moreover, stated that she had no rights to use her personal judgement or impression when judging

students' performances. However, she contended that, "*If it was my own course, I would have changed the criteria. I mean I'd edit them when I find problems*".

Muun

Muun said that she liked the assessment of FE 2 from the first time she used it, yet she realised that there were many teachers who did not like it. She liked it because it allowed a lot of freedom for the learners and the course materials prepared the learners for the assessment. In addition, she pointed out that the assessment of the course encouraged students to have confidence in using English both in writing and speaking forms. Muun added that teachers were the ones who had to give the students the confidence. She maintained that it was important that teachers must not judge the students' performances based on accuracy because if teachers used accuracy to rate students' performances, the students would not speak. She stated that "*in the past that we failed. Students didn't dare to speak English because they were afraid of making mistakes. We only checked accuracy. We didn't check if it was comprehensible. We failed*". Muun stressed that her ideal assessment was "*a series of assessment and authentic assessment*". She believed that a series of assessment, being done continuously, would make assessment as part of learning. She believed that a series of assessments could identify the learning progress of the learners. She emphasised that, "*We don't separate assessment from learning. If we want the learners to benefit the most from assessment, we must have assessment as part of learning ... They have to come together*".

Ronnie

Ronnie believed that there should be various ways to assess students and done many times. He did not believe in the use of only 2 tests to make decisions about students. He said that, "*we can't learn everything at one time and be assessed on it at one time*". Therefore, he proposed that for language assessment, the assessment "*should*

be done continuously in order to see the progress of the learners". Moreover, Ronnie believed that including performance-based assessment in the course was better than having only exams. He considered the results from a performance-based assessment were the indicators of students' ability. Ronnie added that the new assessment helped decrease the stress of the learners.

Ronnie, however, was aware that having many assessment activities could be time consuming especially for scoring. Though each performance-based assessment task did not weigh much (such as 4% for an oral task), Ronnie pointed out that it took a lot of time to rate because there were many aspects he had to check when he scored or rated them, for example the accuracy of language used, required content, body language. Ronnie, moreover, was aware of the drawbacks of the course's assessment such as subjectivity of teachers. He, however, believed that there were enough benefits of the performance-based assessment. Ronnie asserted that, "*If there wasn't any of this assessment, we couldn't really assess students*".

Pawida

Pawida stressed that she had never liked testing with only mid-term and final exams. She believed that they were not motivating. She said because she had background in language education, "*I don't like teaching and learning which depend on one or two tests. I've never liked that kind of assessment. I don't believe in it*". She emphasised that "*I don't believe in assessment as learning … From one exam and made a decision, it was a waste of time*". As one of the team who designed the first two FE courses, Pawida commented that "*I liked them very much because I think about motivation all the time. And when we designed these courses … we thought of motivation*".

Pawida stated that she could be fair within the sections she taught but she pointed out that the teachers in other sections should be fair as well. Since so many teachers were involved in the course, Pawida recognised discrepancies existed in the

different ways teachers rated students' performances. She proposed that standardisation would solve the problem. Pawida also cited that there was a standardisation for the course that semester regarding the written task assignment in which she found that, "*The result was that the majority of teachers gave a score of 3 but I gave only 2.5 [to the given piece of written task assignment] which means that I would have to adjust myself when I rate the following tasks*".

Moreover, Pawida believed that self-assessment was a very important aspect in learning. She determined to have students assess themselves because she wanted them to feel their improvement and believed it would help them learn better. She believed that "*self-assessment is the best assessment*" and it could motivate students to learn. However, she felt that the course had not achieved what she expected. She was very worried with the way students failed to assess their vocabulary knowledge as suggested by the ways they kept their language notebook.

4.2.3.2 Thinking about the written assessment

Arkom

Arkom thought that one of the main strengths of the written assessment was that he could give feedback to students and they would learn from his comments and make improvement. The students had to submit a first draft, though not required by the course, so Arkom could ensure that the topics of each group were different and that students did not copy them from other sources. He added that "*Maybe it's pessimistic but they might have cheated"*. Despite the fact that teachers should not return the written tasks to students (as discussed above), Arkom returned the written tasks to students with feedback. One of his goals in giving feedback was that for students to discuss their tasks with him as so he could explain his comments and reasons behind why they got a particular mark. He stated that, "*For example, I could tell that one student translated from Thai directly to English, which was grammatically incorrect.*

So students would feel that they needed to improve in this area". He also believed that by giving feedback to individual students, they would, consequently, discuss the comments with their peers. In doing so, students would then learn from their peers. He stated that "*most students improved this way, though some didn't*".

Wawan

Though the written tasks were done as process writing, they were all done as homework. This raised Wawan's concern of students' cheating. She stressed that as students did not do their homework and prepare for the class, she could give informal feedback only to the groups who prepared. From her observations, these groups usually got good marks. Regarding the issue of feedback, Wawan agreed that giving feedback was crucial. However, because of limited time in the class period, she could not give feedback to every group. Wawan pointed out that giving feedback to individual groups required a lot of time. Since there was no time in class, in order to do so, it was necessary to make appointments to see the students outside class. Wawan said that giving individual consultation was not in the *Teacher's Guide* nor in the programme. Therefore, she did not do it. What she could do was to walk around during class and try to give some comments like circling mistakes while the students were working. Nevertheless, she admitted that she felt it was not good because "[students] *should get feedback and improve the assignments themselves"*.

Muun

In Muun's opinion, the written tasks did not encourage students to think. She said that students only just "*cut and paste*". They copied from the given examples and made some changes. She did not think that the students' performances from doing the tasks represented their writing ability. She proposed that students had to close the book and write. They could still work in groups, she added. Muun believed that because the tasks were done in groups or pair works, students learned a great deal from those

collaborative activities. She said that, "*They have to help each other. When working in groups, students learn from each other. They learn much more this way; more than from teachers*".

With the problem of students cheating on the written tasks, Muun pointed out that, "*Students only cheat when they aren't confident*". She believed that once the students felt that they could write, they would not cheat. Thus, Muun explained to the students that it did not matter if they made grammatical mistakes in their written assignment. She told them to write in class. Furthermore, Muun pointed out another problem with the written tasks (which is also applicable to the oral tasks) that the criteria were not clear enough, which she believed to be unfair. She also stressed that when assessment was not fair, it was unreliable and invalid.

Ronnie

Ronnie thought that grammar and vocabulary were very important aspects of the written tasks because they were the basis of what conveyed meaning to the readers. Thus, when he rated the students' written performance, Ronnie focused on accuracy of grammar and vocabulary usage. He suggested that one reason the students made a lot of mistakes could be due to the fact that they used direct translations from Thai to English in order to complete their written assignment. Thus, there were many mistakes because direct translations do not always make sense. Ronnie stressed that students should have been aware of what aspects their tasks would be rated because the detailed criteria were explicitly available to them in their *Student's Workbook*.

Pawida

Pawida was not content with the criteria for the written task. She emphasised that the criteria did not cover the quality of the work, but the quantity. She stated that, "*This is a fixed form – a format. No one missed it. Quantity is not important, but the quality. They didn't miss the quantity part ... Most students met those requirements*". She

believed that the domain ‘quality’ would help with the criteria for oral tasks. In addition, Pawida also thought that the criteria were not clear enough. She stated that with the given criteria, it was not possible to check the written tasks in detail. For example, the criteria stated that “*there must be emotive adjectives, comparative adjectives, and factual information*”. Pawida pointed out *'They don’t tell us how many, do they?”* Pawida, moreover, thought that ‘grammar’ should not be part of the criterion domains because she believed that the most important aspects for the task were content, relevancy, comprehension and logic. Another aspect of the criteria in which Pawida was dissatisfied with was that partial marks were not allowed. She exclaimed, “*I was so worried. 3? But it was better than 3*”.

4.2.3.3 Thinking about the oral assessment

Arkom

Arkom pointed out that for a role-play, acting (one of the criterion domains) should not be the focal point of rating. He said that, “*This isn’t a drama class in which students only act and spend money on props. And if they rehearse well, they get good mark … This is a language classroom*”. Thus, Arkom proposed that use of language should be the main focus. Yet, because the task was in a role-play form, students needed to practice. Likewise, as mentioned above, Arkom believed that content (one of the criterion domains) changed all the time; therefore, it should not be the focus when teachers rated students’ oral performances. Another domain in the criteria that Arkom disagree with was ‘creativity’, for the same reason that he did not agree with the domains of ‘acting’ and ‘content”’

Wawan

Ideally, Wawan wished to focus on communicative ability when she assessed students. She wanted to check if the students could communicate, especially what they had learned from the course, with other people in English or with foreigners. She

pointed out that using tests or exams was not enough. She wanted to have each student speak. Moreover, the content of what the students said must be original.

Muun

In Muun's opinion, fluency was very important for oral assessment because it signified the level the students' communicative ability in terms of their skills. In addition, Muun believed that there were two kinds of fluency: from the brain and from a rote-memory. She was aware that every student tried to memorise the scripts. However, Muun was aware of the importance of rote-memory, but teachers had to move students forward. Muun stressed that if there was only rote-memory, the presentation would not be natural. This case could be seen in the role-plays, she added. Going from pure memorisation, the students could not make the role-play go on smoothly when they encountered problems. The role-play would be dead. Muun cited that, "*There were times when everything stopped – they couldn't continue*". She also noted that this happened because the students did not risk using different language and they lacked the confidence.

Ronnie

Ronnie was worried that he might not be fair and accurate when he rated students' oral performances because of the limited time for each presentation. His solution was following the criteria as strictly as possible. In regard to rating, Ronnie was concerned with the difference in rating between Thai and native speaker teachers. Being an assistant coordinator of the course and one of the material writers, Ronnie reported that he found the native speaker teachers tended to give higher marks. Ronnie pointed out that these teachers did not pay much attention to other aspects of the criteria. He questioned if these teachers were equally strict when rating students' assignments. Furthermore, Ronnie admitted that he did not award any student a full mark for their oral presentations. For him, a full mark equalled perfection. He said that students

made mistakes in both sentence construction and vocabulary. Students mispronunciation of the key words led to a distortion of meanings, making it difficult to comprehend. Their acting was not great either, he added. However, he said that a full mark was possible "*if the presentations were well prepared*".

Pawida

Pawida emphasised that she liked oral assessment. As one of the FE material developers and writers, Pawida said that having students do oral presentations was one of the main objectives in revising the new courses. However, similarly to her views toward written assessment, Pawida disliked the criteria. She thought that the criteria were not clear enough. She said "*What does 'correct' mean? No mistakes and correct use of grammar?*" Pawida also disagreed that grammar should be one of the criterion domains because she thought it was very difficult to rate grammar on spoken discourse. Like written assessment, Pawida also thought that the criteria for oral tasks should include quality of the task requirement, not only quantity. She believed that a 'quality' aspect of the task could distinguish between good and weaker students.

4.2.3.4 Thinking about the final examination

Arkom

Arkom thought that the exam was based on the objectives of the course and the tasks, and was not too easy or too difficult. Therefore, he thought that the exam was a good assessment of students. However, there were some problems with the exam, especially its format and layout. Students had to write on page 8 of the exam paper but the reading part was on page 7. Arkom noticed that some students wrote the answers on a different sheet of paper, then copied them onto the answer page which was a waste of time. He also did not agree with the weighting of the exam items. In the reading part, Arkom thought that the items on references were easier than guessing the meanings of unknown word items. He suggested that they should weigh

only half a point whereas guessing meaning should weigh 1 point. Moreover, he thought that the exam should weigh only 50 instead of 56%.

Wawan

Wawan agreed that the exam results of her students correlated with their performances in class. It also reflected how the students performed in general in class. When the students failed to use what they learned in class, she said that "*This shows that they didn't use the strategies we taught in class. We taught them how to guess meaning of unknown words from context ... But they didn't use it*". Since the exam weighted 56% of the course assessment, Wawan said that she could guess the grades of the students from the exam results. Moreover, she agreed that cloze test was difficult because it required students to understand the reading, know the meaning of the words, and understand the grammar in order to fill in the blanks. Thus, she tried to help the students by giving them a guideline: "*For example, I told them that if they saw 'is', the following word could be adjectives*". However, she did not have any problem marking the exam because the guidelines and answer keys were very clear.

Muun

Muun pointed out that the exam contained varieties of item types – multiple choice, True/ False, note-taking – which test different aspects of students' ability. She added that the exam had high discrimination power. She said that the exam "*has to be able to assess [discriminate]. For example, weaker students should fail but good students should do well. I think this is a good characteristic of a test"*. Moreover, Muun stated that she, especially, liked the writing part because "*it reflects the relationship between classroom and examination conditions. If students can do this part well, it means that they'll be successful. But if they don't do well, they'll fail*". However, she thought that the exam was quite difficult especially the cloze test. She stressed that the students didn't do well on cloze because it was difficult. Furthermore, Muun noted that one

major weakness of the exam was a change in the criteria for scoring the writing part after the exam. She emphasised that this was unfair for the students and also unethical. In addition, when she realised that there was a change, she had to re-mark all the papers.

Ronnie

Ronnie, as one of the exam item writers, was satisfied with the exam in general because he agreed that it was based on the objectives of the course. However, he was not satisfied with the answer keys. He said that there was more than one possible answer to some items in the grammar part. As part of the exam committee, Ronnie admitted that during the meeting, they did not pay enough attention to the grammar part. Ronnie said that "*we didn't look at the grammar part often enough, so there were some mistakes"*. Furthermore, Ronnie agreed that the reading of the second passage was comparatively difficult, thus students might have more problems trying to understand it. He thought that those items, which only weighted half a point, should weight one point instead. He added that the number of items did not have to be so numerous and there should be more 'given' answers in the order-filling part because if students gave only one wrong answer, they might lose all points for that part. He stated that, "*There was a chance some students might misunderstand the reading and get all answers wrong"*. In addition, he was not satisfied with the guessing meaning part because the item type was not tested the same way as it was taught in class. Ronnie, however, was satisfied with the cloze section. He thought it was not too difficult. He believed that cloze tests could test the ability of students because if they really understood the passage, they could choose the right words.

Pawida

Pawida thought that the exam contained too few questions. She said, "*I felt that 2 questions weren't enough for this part … The reading passages are long but there*

were too few questions". In addition, some items were problematic as some questions relied on the correct answers of previous questions, some items asked the same thing, some items only weight half a point and the context clues for some items were too obvious. Moreover, she pointed out that the exam questions did not encourage the students to think. She suggested that, "*We could have this question [asking for opinion] as part of a comprehension question, but we have to give credit for their opinion as well*". From another perspective, Pawida agreed that the exam tested the students' ability and what was taught in class. Nevertheless, she argued that it could be done through rote-memorisation. She thought that the writing part, the itinerary writing, was a repetition of what was done in class. Pawida also thought that the exam should weigh 40%. She thought 'it's too quick for 56%". She also thought that the in-class exercises did not prepare students well for the exam because the practice review was far too easy. She suggested that the exercises should include more items and be of similar length and level of difficulty as the actual exam. She said, "*Look here [at the in-class exercises] – now look at the exam. The exam contains paragraph level questions but the quizzes are instead discrete points. They're different"*. Moreover, she asserted that the answer keys had to be very clear and accurate.

4.2.4.5 Reported practices in assessment

Arkom

Arkom stated that the course's assessment was flexible. Though Arkom said that he followed the criteria when he rated the students' performances, he added what he thought fit depending on the section and used different standards when rating students from different sections. He said that because he knew his students he "*rated students' performances using the standards of that particular class*". In addition, he gave higher marks for written tasks than the oral tasks. As mentioned above, Arkom believed that the target of the course was oral performance, so he was especially strict

when it came to oral assessment. Moreover, he applied the criteria from different courses to the FE courses. For example, he deducted points for responsibility, which was not included in the criteria of the tasks. When one group did not submit their first draft, which was not required by the course, Arkom deducted points from this performance. Moreover, as mentioned above, he believed that content and creativity were relative, so he gave these two domains a full mark but focused on the domain 'language". He said that it was not possible to objectively evaluate these aspects. Thus, he decided to focus on the use of language when he assessed the students' tasks. Arkom reported that he tended to give students low marks for their performances because he wanted to push students to work harder for the final exam.

Wawan

For written assessment, Wawan thought that the given criteria were too detailed and hard to follow. Thus, she used the criteria as guidelines while adding her own judgement when she rated the students' performance. When Wawan rated the students' written and oral performances, she did not give any student a full mark. She said that "*I didn't want to give a full mark because there were mistakes. It was impossible that anyone would make no mistakes. - - The criteria stated 'correct use of grammar and spelling', but no one had 'correct'"*. Despite the fact that the instructions for the oral tasks stated that teachers must award every member of the group or pair the same mark, Wawan did not follow it. She explained that she had discussed this issue with other teachers and they agreed it wasn't possible to award every member of the group the same mark. She pointed out that it was not right to give a student the mark which was not their level of proficiency. She said, "*We know that students can't reach that level and so we can't award them for that level"*. In addition she thought it was not fair for other students in a group who had higher proficiency level but got the same mark as the lower proficiency students. Wawan justified that the results from her ratings correlated with the final exam results. She

added that, "*We can't use the marks from classroom assessment to help students when they can't do well on the final exam. It will make them even weaker in the next course*".

Muun

Because Muun believed in a series of assessment and integrated assessment into many aspects of her classroom teaching. In her class she used a variety of assessment that students do not notice or feel any differences between the ordinary teaching days or assessment days. In addition, Muun was very flexible with the dates for assessment. She added that, "*When the students were not ready for the oral tasks, for example, during the week with assessments, I'd postpone the assessment date for them"*. Moreover, Muun believed that every student had the potential to learn but it depended on their motivation. She said that it was the responsibility of a teacher, especially for language teachers, to create this motivation. She also believed that it was the teachers' responsibility to give students moral support, which was related to the students' motivation and self-study. She explained that she offered the students moral support by giving them verbal compliments when they did well on their oral presentations. Consequently, the students would feel that they were important and thus motivated to do self-study outside class. Another way Muun gave students support was by rewarding them. Citing existing teaching theories, Muun pointed out that teachers must use a reward system to reinforce the students' learning. She admitted that she awarded quite high marks for her students on the tasks because she believed that it was a reward for students.

Ronnie

Ronnie emphasised that when he rated the students' performance, he always followed the criteria. Because he was aware of the possible subjectivity caused by performance-based assessment, he argued that the criteria helped controlled how he

rated the students' performances. However, Ronnie admitted that when he marked the students' final exam – because of some personal matters he needed to finish marking the exam quickly – he deviated from what he usually did for the writing part. He used the criteria, which were analytic criteria, to give a holistic score. However, he justified that "*I was in a hurry but I followed the criteria … I didn't just read through and assign the marks. I checked different aspects. I didn't use my impression. I followed the criteria*".

Pawida

Pawida usually gave feedback on students' oral presentations the following period after presentations. She explained, "*I would tell them how they did on their oral presentations"*. In addition, when she wanted to give individual students her comments, she would talk to that student privately to spare them any potential embarrassment in front of their peers. Moreover, when she wanted to give overall comments, she would write the common mistakes on the transparencies and show them on the OHP or write them on the board. However, Pawida found that she could not do everything she wanted because of the norms. She stressed that because there were so many teachers teaching the course, everyone had to use the same set of standards in order to achieve fairness. Though Pawida did not like the criteria for both written and oral tasks, she reported that "*I followed the criteria even though I didn't agree with them – for the sake of standards. I didn't agree with them*". She later realised, though, that the criteria did not state very clearly about how to rate grammar. Therefore, Pawida paid more attention to the communicative aspects of the tasks. She said of her students, "*They could communicate very well. They could write scripts – I was very happy, so I gave them quite high marks*".

4.2.4 Discussion

From the overview of the data from the five teachers and assessment practices in the Department presented in the previous section, the most prominent problem in assessment in the Department is how teachers view and apply the rating criteria. The course required teachers to use the same materials and assessment. However, the findings, as a reflection in their beliefs, attitudes, and understanding of language assessment, revealed that each teacher had different views towards the rating criteria and used them differently.

4.2.4.1 Different views toward rating criteria

With performance-based assessment, rating criteria are one of the most important components of the rating process. The criteria used in these two foundation courses have been revised many times. However, drawing from the interview data, the criteria did not seem to meet the expectations of the teachers because they were not clear enough. The present criteria are also controversial because some teachers thought that they were too detailed whereas others thought they included insufficient detail. According to Muun, "*the criteria create discrepancies in teachers' judgments' because they do not clarify what each level of descriptors mean*". She added that 'the criteria must tell us what the score "4" means in order to achieve reliability and validity.' Pawida had a similar view with Muun. She said that she had problems with following the criteria when she rated students' performances. The criteria were not clear and detailed enough for her, especially with the oral assessments. Pawida also admitted she was not very happy with the ways she had to rate students' performances in the course. Though she was not satisfied with the criteria, she had to follow them. She said that she could not use her own impression or create her own criteria because it would not be fair for her students and other students in other sections. From my observations, the criteria of the tasks or final exams were

occasionally changed after the tasks had been submitted or students had already taken the exam.

Though Wawan agreed to a certain extent with Muun and Pawida that the criteria were not clear, she thought that they were too detailed. She pointed out that she had problems trying to follow the criteria when rating students' performances. Arkom, on the other hand, did not have any problems with the criteria. He said, "*I think that the criteria are good enough. They have been piloted and revised. I feel that they have been created with the best efforts*". Similarly, Ronnie, as one of the course material writers who designed the rating criteria, was satisfied with the criteria. He urged that every teacher should follow the criteria to solve the problems with reliability.

4.2.4.2 Different applications of rating criteria

Muun reported that she used her impression with an emphasis on the communicativeness of the tasks when rating students' oral presentations. From my observations, she rated the performances holistically despite the fact that the rating criteria were analytic. She said, "*I checked the overall performance of each group. This level of impression was 1 point, this level, 2*". She admitted awareness that her way of grading might be different from other teachers. Likewise, Pawida also emphasised communicative aspects, regardless of the criteria, when she rated students' performances, especially on oral assessments. She said, "*I'm very happy if the students can communicate. If they can communicate, I give them a full mark. Grammar isn't the most important aspect, but comprehension*".

In contrast, Wawan paid a great deal of attention to the accuracy of language used when she rated students' tasks. She did not give any student a full mark because of incorrect use of grammar usage and spelling. Wawan, however, had a different way of using the criteria. Her way of treating the criteria for oral performance was different from that of written performance. For the oral tasks, she added her own

judgment when she rated student's performances. She said "*I had to use my own judgment and consideration apart from the given criteria*". Nevertheless, she followed the criteria strictly for the written tasks despite her frustration. She admitted that "*I had to follow the criteria though I didn't quite agree*". Ronnie, similar to Wawan, did not give any student a full mark. He said that in order to get a full mark, students' performances had to be "*perfect*", especially for oral tasks. In addition, Ronnie emphasised that he followed the criteria very strictly when he rated students' performances. In order to achieve fairness, Ronnie pointed out that all teachers had to follow the given criteria strictly. He said, "*I want all teachers to use the same standards by following the criteria strictly and trying to eliminate their own impression or subjectivity"*.

Arkom, on the other hand, had a different way of handling the criteria. He admitted that he followed the criteria but also added what he thought important in different circumstances. He paid attention to different domains of the criteria for different tasks. Moreover, Arkom treated oral tasks and written tasks differently.

4.2.4.3 Insufficient understanding in language assessment

From my observations, it seems that teachers' lack of knowledge in the area of language testing and assessment clearly affected the ways in which teachers dealt with assessment. There are evidences indicating that teachers in the Department lacked sufficient understanding of basic concepts in assessment. For example, during the study, one teacher asked me whether it was important to know the definitions of technical terms in language testing and assessment. For him, he would rather pay them no attention because he thought that as a teacher who rarely got involved in test development, he did not need to know them. In addition, the majority of teachers were not aware of the concept of assessment for learning or formative assessment. For example, one teacher told me that the only way to do formative assessment was to have students do formal assessment many times during the semester. Another

important piece of evidence illustrating teachers' lack of understanding about the core concept of language testing and assessment was how the Department conducted the standardisation (see Section 4.2.2.2). Finally, the decisions in the Department grade meeting (see Section 4.2.2.3) were based heavily on the conventions which have been passed down from the previous years, not on assessment theories.

Because teachers do not have sufficient background, and because of the misleading project standardisation procedure, they had very different ways of handling the assessment. The concrete evidence illustrating this point in my observations is that some teachers did not check students' attendance every class period. Since the course requires teachers to deduct 2% for each absence, it is not fair for students in other sections in which teachers always checked attendance. The most important consequence caused by the lack of effective and sufficient background in language assessment and rater training is that different teachers interpreted and used the criteria and rating scale differently, as indicated in the findings from the case studies described above. Some teachers, for example, used the criteria as a guideline when they graded students' performances, some used their own impressions, or some used their own judgments but added what they wanted, whereas other teachers followed the criteria strictly.

4.3 Conclusion and Implications for the Main Study

The findings from the pilot study support McDonough and Chaikitmongkol's (2007) conclusion that teachers have positive reactions to including of performance-based assessment in the FE 1. However, from the analysis of the data, there are considerable differences among the teachers' views toward the rating criteria and how they are applied. For instance, they disagreed greatly on the importance of accuracy in rating students' oral performances, and they interpreted and used the domain 'accuracy' of

the criteria differently. In other words, the teachers had mistaken views of how to conduct performance-based assessment and therefore had done it inappropriately.

If teachers in this context did not have appropriate understanding of the rating process and its related issues, and could not agree upon the rating criteria, it would impose serious problems on the quality of assessment at the Department as a whole. It seems that a lack of in-service professional training as well as differences in teachers' educational background, teaching experience, assessing, material and test development experiences are responsible for these differences. They also affect the teachers' competence and confidence, and the way they conduct assessment in their classrooms. As described in Section 4.2 that teachers in Thailand have very limited education in language testing and assessment along with more complex assessment systems required by the educational reform, I believe that to solve the problems in this context is to implement an in-service professional development programme in language assessment for the teachers. This professional development programme would focus on getting teachers to cooperate, in a series of workshops, to evaluate the existing rating criteria, and to create and evaluate a new set of rating criteria [if required]. These activities would aim to provide the teachers with a practical and theoretical understanding of performance-based language assessment so they could become more competent and confident in conducting assessment activities in their classrooms and participate in assessment activities in the Department.

The focus of my research, therefore, became the development of teachers who participated in the professional development programme. I used the insights I had gained from the pilot study in designing the main study. In the following chapter (Chapter 5), I describe the research methodology, research process and data analysis for the main study and the follow-up study. The data from the main study will be described in Chapters 6 and the follow-up study in Chapter 7.

5 Research Methodology, Research Process and Data Analysis

This research project is a longitudinal qualitative study conducted in three phases: pilot study, main study, and follow-up study. This chapter explains the research methodology underlying the research project, its research process and data analysis. In the first part, I explain the theoretical principles of the research methodology and methods utilised within the main study and follow-up study. (The pilot study is reported in Section 4.2.) The second part includes the processes of these studies, consisting of the purposes and research questions, data collection process, and participant profiles. The final section describes the analysis process of the data.

5.1 Research Methodology

From the beginning of the study, I have employed different research methodology stand-points in my data collection activities and analysis. In the pilot study I used both qualitative and quantitative approaches (see Section 4.2.1.3). However, from data analysis, it became apparent to me that a qualitative approach was the most appropriate for this study in this particular context to answer my main research questions. Therefore, for the main study and follow-up study, I only employed qualitative research methodology. Apart from investigating the thinking and experiences of individual teachers, another major aim of the present study was to examine the social system as well as the interactions of the teachers within their social environment; therefore, for the main study and follow-up study, I only employed qualitative research methodology. This methodology was selected because it offers data collection methods which potentially enable the researcher to gather rich data for the stated purposes. In addition, this methodology provides methods of analysis that are grounded in the data itself. Furthermore, the data collections were

planned to be carried out at different points and over a long period of time in order to describe changes of individual teachers, and qualitative research methodology lends itself to this plan. Qualitative methodology will be discussed in detail in the following sections.

First of all, it is important to make the distinctions between the terms methodology and method. Cohen, Manion and Morrison (2007) define the term ‘methods’ as a 'range of approaches used in educational research to gather data which are to be used as a basis for inference and interpretation, for explanation and prediction. The term ‘methodology’, on the other hand, is used to 'describe approaches to, kinds and paradigms of research' (p. 47). Drawing from these definitions, the research methodology I have employed as the major part of my study can be categorised as a qualitative approach which includes multi data collection methods. In the following sub-sections I elaborate on the definitions and characteristics of the qualitative research approaches I have adopted in the study and the methods employed in the data collection.

5.1.1 Qualitative research

Hesse-Biber and Leavy (2006) identify the epistemology of qualitative research as a hermeneutic or interpretive perspective which is based on the interpretation of interactions and the social meaning that people assign to their interactions. Qualitative researchers are interested in generating theory, relying heavily on ‘inductive models’ where the theory develops directly out of the data. They often employ more than one of the following methods within the context of one research project to develop larger theories about social life that emerge from the people who experience the aspect of social reality being studied (though this is not an exclusive list): ethnography, in-depth interviewing, oral history, focus group interviewing, case study, discourse analysis, and content analysis (p. 9). However, Hesse-Biber and Leavy warn that ‘multimethod designs, in their best execution, do not simply rely on

more than one method of data collection for the sake of yielding 'more data' per se. When multiple methods are used, the methods interact with each other and inform the research process as a whole' (p. 20).

Different scholars have categorised approaches in qualitative inquiry differently. For the purpose of this thesis, I adopt the framework put forward by Creswell (2007) who identifies five major approaches within the inquiry: narrative research, phenomenology, grounded theory, ethnography, and case study. In the study, I employed three approaches: grounded theory, ethnography and case study. In addition, longitudinal research design and action research were incorporated in the design of the study.

5.1.1.1 Grounded theory

Grounded Theory (GT) is a strategy of inquiry, consisting of a set of data collection and analytic procedures, in which the researcher derives a general, abstract theory of a process, action, or interaction grounded in the views of the participants (Charmaz, 2004; Creswell, 2009). GT methods allow researchers to conduct qualitative research 'efficiently' and 'effectively' because these methods provide systematic procedures for shaping and handling rich qualitative materials (Charmaz, 2004, p. 497). Charmaz (2002) points out that grounded theory consist of guidelines that help researchers to study social and social psychological processes, direct data collection, manage data analysis, and develop an abstract theoretical framework that explains the studies' process (p. 675). The methodology was first introduced by Glaser and Strauss (1967) in *The Discovery of Grounded Theory.* They, for the first time, made explicit the analytic procedures and research strategies that previously had remained implicit among qualitative researchers. Since then, GT has developed in many directions. Dey (2004, p. 80) emphasises that:

> there is no such thing as ‘grounded theory’ if we mean by that a single, unified methodology, tightly defined and clearly specified. Instead, we have different interpretations of grounded theory – the early version or the late, and the versions according to Glaser (1987), or Strauss (1987), or Strauss and Corbin (1990), among others (e.g. Charmaz, 1990; Kools et al., 1996).

Nonetheless, Charmaz (2002, p. 677) points out that all variants of GT share the following characteristics: simultaneous data collection and analysis, pursuit of emergent themes through early data analysis, discovery of basic social processes within the data, inductive construction of abstract categories that explain and synthesize these processes, sampling to refine the categories through comparative processes, and integration of categories into a theoretical framework that specifies causes, conditions, and consequences of the studied process.

It has been noted by many scholars that GT has been widely adopted by researchers in the fields of nursing, education, and many other disciplines. Miller and Fredericks (1999) state that GT can be used to ‘direct the research process as well as provide a heuristic for data analysis and interpretation. In the field of Teachers of English to Speakers of Other Languages (TESOL), GT has a strong appeal to practitioners in the field because it offers a means of developing an understanding of an educational context without demanding the extended exposure for a full ethnography (Richard, 2003, p. 17). Because the present study aimed to examine an educational context and psychological development its teachers, GT was adopted as a means to direct data collection as well as data analysis. In Section 5.3.1, I explain how GT can be used for the data analysis and interpretation based on the latest work by Corbin and Strauss (2008).

5.1.1.2 Ethnography

Ethnography is a strategy of inquiry in which the researcher studies an intact cultural group in a natural setting over a prolonged period of time by collecting, primarily,

observational and interview data with the aim of getting an in-depth understanding of how individuals in different cultures and subcultures make sense of their lived reality (Creswell, 2009; Hesse-Biber & Leavy , 2006). Ethnographic studies intend to explore the culture or shared experiences by understanding the attitudes, knowledge, beliefs that influence the behaviours of the people within a community (Lodico, Spaulding & Voegtle, 2006). Ethnographers depend on 'key informants' in providing them with the 'richest insights into the culture', 'the issues addressed in the study' as well as the 'unwritten rules' of the group. Moreover, ethnographic reports usually comprise a 'thick description' of the situation 'capturing the full complexity of the nuances in interactions, cultural practices, and beliefs of the group under study' (p. 268). However, the participants or key informants might show and tell what they think researchers want to see and hear, as well as hide things and tell lies (Delamont, 2004, p. 212).

In the field of TESOL, an ethnography could be a study of a group of teachers in their institution over a term or year in which the researcher could join the staff as a temporary teacher in order to take field notes, observe classes, interview teachers, and record some staff meetings, for example (Richards, 2003, p. 15). Richard (ibid.) also suggests that ethnography provides a means of understanding teachers' own professional worlds. In the present study, though I was not officially in the field as a member of staff, I was recognised as a member of the Department who was on study-leave. In the field work, I attended and recorded staff meetings and frequented the Department on a regular basis (for further detail of my roles as a researcher, see Section 5.2.4; see also ethnography observation, Section 5.1.2.3).

5.1.1.3 Case study

A case study can be defined as a strategy of inquiry in which the researcher explores in depth one or more individual, a programme, process, event, or activity (Creswell, 2009, p. 13). Based on Stake (1995), Creswell (ibid.) describes cases as being

'bounded by time and activity, and researchers collect detailed information using a variety of data collection procedures over a sustained period of time'. In ethnographic research, case study could provide researchers with a thick description of the situation to 'capture the full complexity and uniqueness of the case information' (Lodico, Spaulding & Voegtle, 2007, p. 270). In the area of applied linguistics, a case study is usually associated with qualitative research in which the case has been the individual language teacher, learner, speaker, or writer (Duff, 2008). In addition, the components studied in the case study approach have been the study of individuals and their attributes, performance, development, and knowledge (Duff, ibid. p. 35). Duff also states that case studies can yield 'a high degree of completeness, depth of analysis, and readability'. They could also generate 'new hypotheses, models, and understanding about the nature of language learning or other process' (p. 43). In the present research, the focus was on investigating five individual teachers' knowledge about language assessment, their beliefs about it and attitudes towards it. Therefore, a case study approach was employed because it could provide an in-depth, complex, and thick description of each teacher.

5.1.1.4 Longitudinal study

Apart from being qualitative, the study is also longitudinal in nature. Thomson, Plumridge and Holland (2003) recognise longitudinal qualitative research as a 'promising new methodology' which is yet 'taking place without a relevant literature to inform and debate the epistemological or practical decisions [they] were making' (p. 185). The main purposes of longitudinal research are 'to describe change, and to explain causal relationships' Dörnyei (2007, p. 79). According to Dörnyei (ibid.) the longitudinal design employed in the present study can be classified as a 'prospective longitudinal study', of which data are gathered at different points in time from the same participants. This type of longitudinal research design was utilised because,

according to Dörnyei, it offers a complex and true reflection life story of an individual participant, which was the major aim of the present study.

5.1.1.5 Action research

Finally, another aspect of research design of the present study is action research. According to Burns (1999), action research focuses on concrete and practical issues of immediate concern to particular social groups or communities and it is usually conducted by and with members of the communities. Mackey and Gass (2005) emphasise that action research is usually initiated from a question or problem. It is followed by gathering data and then analysing as well as interpreting the data. Mackey and Gass add that a solution to the research problem might emerge from the findings. The final step of action research could be disseminating of the findings. In addition, Mackey and Gass point out that a change to current practice could be one of the outcomes of action research. Furthermore, Burns (op. cit., p. 35) perceives the process of action research as a series of interrelated experiences involving the following phases: exploring, identifying, planning, collecting data, analysing/ reflecting, hypothesising/ speculating, intervening, observing, reporting, writing, and presenting.

In terms of locating action research in the research paradigms, Burns (2005, p. 61) proposes the following characteristics of action research:

Philosophical assumption: People within social situations can solve problems through self-study and intervention

Purpose: To develop solutions to problems identified within one's own social environment

Main methods: Mainly qualitative, interpretative, cases studies reflectively through cyclical observational and non-observational means

Outcome: Action to effect change and improvement, and deeper understanding in one's own social situation

Criteria for judgement: Subjectivity, feasibility, trustworthiness, and resonance of research outcomes with those in the same or similar social situation

From extensive review, Burns (ibid., p. 62) identifies the following purposes and scope of action research activities in the field of language teaching:

- to address and find solutions to particular problems in a specific teaching or learning situation,
- to underpin and investigate curriculum change or innovation and to understand the processes that occur as part of an educational change,
- to provide a vehicle for reducing the gaps between academic research findings and practical applications in the classroom,
- to facilitate the professional development of reflective teachers,
- to acquaint teachers with research skills and to enhance their knowledge of conducting research, and
- to enhance the development of teachers' personal practical theories.

Moreover, according to Burns (2009, p. 292 - 293), action research can be grouped into three categories:

1 Required components in formal undergraduate or postgraduate courses – in which teachers typically undertake small-scale projects that results in term papers, class presentations, or PhD dissertations

2 Collaborative teacher-researcher projects within educational organisations/ programme – which involves teachers in large-scale institutional curriculum change and continuing professional renewal

3 Individual projects by classroom teachers/ teacher educators

In the present study, the PD programme can be viewed as an action research in which five teachers collaboratively conducted a research project with the researcher to solve problems in assessment embedded in the Department. Apart from solving these problems, another primary aim of the programme was to provide these

teachers with theoretical and practical understandings of performance-based language assessment. At the same time, the researcher collected qualitative data to investigate the impact of the programme on these teachers.

5.1.2 Data collection methods

From reviewing a large number of the studies in ESL/EFL teacher cognition (including teachers' beliefs, knowledge, attitudes and practices), Borg (2006) reports four most widely used methods in these studies, which consist of:

- *Observation* – including structured, unstructured observations of classroom practices;
- *Self-report instruments* – including questionnaires, scenario-rating tasks, and test;
- *Verbal commentaries* – including structured, scenario-based, repertory grid, semi-structured, stimulated recall, think-aloud protocol; and
- *Reflective writing* – including journals, autobiography, retrospective accounts, and concept mapping.

Furthermore, Borg comments that each method has its own advantages and disadvantages; thus, multi-method strategies, or combining methods, have been adopted by the range of studies. He also reminds us that in selecting data collection methods and making claims, it is very crucial to be aware of the underlying assumptions about teacher cognitions reflecting from different kinds of evidence. For instance, it is implied in *self-report instruments* that 'beliefs can be articulated and rated against predefined propositional statements and understood without direct reference to actual instructional practices' whereas in *interviews* 'beliefs can be articulated orally and that teachers are able to provide a verbal account of the cognitions underpinning their work' (p. 279).

Sakui and Gaies (2003) also add that the studies of teachers' beliefs should employ different methods such as interview and observation data, diary and journal entries, and surveys. They believe that the qualitative data has helped clarify the

relationships between teacher cognition and context factors and the situated nature of teacher cognition. Although I employed different methods in the course of this research project, in this thesis I only report in detail the methods contributed to the main discussions, including: interviews, focus groups, ethnography observations, and think-aloud.

5.1.2.1 Interviews

In general 'an interview is a conversation, usually between two people… where one person – the interviewer – is seeking responses for a particular purpose from the other person: the interviewee' (Gillham, 2000, p. 1). Rapley (2004) adds that in qualitative research, interviews are 'social encounters where speakers collaborate in producing retrospective (and prospective) *accounts* or *versions* of their past (or future) actions, experiences, feelings and thoughts' (p. 16, emphasis in original). Moreover, an in-depth interview, which usually consists of open, direct, verbal questions, is the methods used when 'the focus of inquiry is narrow, … the respondents are familiar and comfortable with the interview as a means of communication, and the goal is to generate themes and narratives (Miller and Crabtree, 2004, p. 189). In the same vein, Hesse-Biber and Leavy (2006) note that this method is useful when the researcher has a particular topic he or she wants to focus on and gain information about from individuals' (p. 120). They also stress that in-depth interviews are 'a meaning-making' and 'knowledge-producing conversation' that occurs between two parties. In applied linguistics research, in this type of interview, while the researcher tries to ask each interviewee a certain set of prepared questions, he or she also 'allows the conversation to flow more naturally, making room for the conversation to go in new and unexpected directions' (Hesse-Biber and Leavy, ibid., pp. 125). Furthermore, Dörnyei (2007, p. 136) states that most interviews conducted are the semi-structured interview and it is suitable when:

> the researcher has a good enough overview of the phenomenon or domain in question and is able to develop broad questions about the topic in advance but does not want to use ready-made response categories that would limit the depth and breadth of the respondent's story.

However, Charmaz (2002) argues that when an interviewer relies on one-shot interviewing, he or she could miss opportunities to 'correct earlier errors and omissions and to construct a denser, more complex analysis'. Therefore, she recommends, especially for GT study, 'multiple sequential interview' as a solution because it could chart a person's path through a process, fosters trust between interviewer and interviewee, which allows the interviewer to get closer to the studied phenomenon and permits independent checks over time. A multiple sequential interview also allows the participant's story to gain depth, detail, and resonance, prompts a fuller story, and allows the researcher to hear about events when participants are in the middle of them, not only long afterward (p. 682).

In addition to the multiple sequential interview method used in the present study, I employed a retrospective stimulated recall technique in certain interviews (see Section 5.2.2.3). This technique allows the researchers to explore the participants' thought process after they have performed a task or participated in an event. The participants are asked to recall and then verbalise their thoughts with support from some sort of stimulus, for example listening to a recording of the participant's own teaching, or showing the person a written work that he or she has produced (Mackey & Gass, 2005; Dörnyei, 2007). Realising the potential problems related to issues of memory and retrieval, timing, and instructions, Mackey and Gass (2005) provide the following recommendations: data should be collected as soon as possible after the event that is the focus of the recall, the stimulus should be as strong as possible to activate memory structures, the participants should be minimally trained, and the level of structure involved in the recall procedure is strongly related to the research question (pp. 78 - 79).

In the present study, though I had interview schedules (see Appendix K), the interview sessions were done very much like conversation. I asked follow-up questions, but the questions used were open-ended because, following GT, I wanted the participants to reveal their own life experiences. However, the prepared questions would help to narrow down to those on assessment and teaching experiences. In some of the interviews, I used stimuli to help participants retrospect what went on in their minds while doing such activities, which included the materials from the course relating to assessment, their ratings of student's written tasks and the materials from the PD (see also Section 5.2.2.3). In the main study, I interviewed the participants three times: before the implementation of the PD, during the PD, and after the PD. For the follow-up study, I interviewed the participant twice: before the implementation of the assessment criteria, and after the implementation of the criteria.

5.1.2.2 Focus groups

A focus group is a research method used to collect data through a group interaction on a topic determined by the researchers (Hesse-Biber & Leavy, 2006; Morgan, 2004). Dörnyei (2007, p. 144) recognises the format of a focus group is generated from 'the collective experience of group'. In a focus group, the participants think together, inspire and challenge each other, and react to the emerging issues and points. Moreover, he points out that the 'within-group interaction can yield high-quality data as it can create a synergistic environment that results in a deep and insightful discussion' (ibid.). Individual interviews might put a great deal of pressure on the relation between the interviewer and interviewee; a focus group, on the other hand, 'can provide prompts to talk, correcting or responding to others, and a plausible audience for that talk that is not just the researcher. So focus groups work best for topics people could talk about to each other in their everyday lives – but don't' (Macnaghten & Myers, 2004, p. 65).

In the present study, two focus group interviews were employed. The first one was conducted in the first PD workshop, after the first round of individual interviews, to encourage the participants to begin sharing their ideas and experiences in teaching in and assessing of the course in a group setting. The second one was conducted in the final PD workshop, before the final round of individual interviews. This second focus group aimed to allow the participants to reflect on the PD and their views toward the assessment of the course.

5.1.2.3 Ethnography observation

In an ethnographic study (see also Section 5.1.1.2), 'participant observation' is a major research method used. This method requires the researcher to 'live or make extensive visits to the setting they are studying, observing as well as participating in the activities of those they are researching'(Hesse-Biber & Leavy, 2006, p. 230). When the researcher is in the research setting, there are different degrees to which he or she participates in the field. The researcher could be a complete observer, observer-as-participant, participant-as-observer, to complete participant. However, in this respect, Hesse-Biber and Leavy argue that there are 'degrees of participation' in the research setting and degrees to which members of the setting view the researcher as an insider of the setting (p. 250).

In the present study, I was in the research setting as a 'participant-as-observer' in which I participated 'fully in the ongoing activities' and my identity was known to the members of the setting that I was conducting a PhD research project on language assessment. Furthermore, because I was perceived as one of the staff members - as I was officially a staff member on study leave and I would return to work there when I finish my studies, I had the privilege of gaining the rapport with the teachers in the Department. I was also allowed to attend all the meetings I requested.

5.1.2.4 Think-aloud protocol

Think-aloud protocols or online tasks require a participant to verbalise 'what is going on through their minds as they are solving a problem or completing a task' (Mackey & Gass, 2005, p. 79). Dörnyei (2007), based on the discussion by Ericsson (2002), reports that this method involves the 'concurrent vocalization of one's "inner's speech" without offering any analysis or explanation'. He also points out that the method 'is not a natural process' and thus 'participants need precise instructions and some training before they can be expected to produce useful data'. In addition, the researchers employing this method need to provide participants with preparation for the tasks (p. 148). In language assessment, think-aloud protocol method has been used widely, for example, Cumming et al. (2001, 2002) and Lumley (2000) used the method in investigating scoring decision while rating writing tasks.

In the present study, the participants were asked to provide think-aloud protocols three times. Before the first think-aloud session, I provided them with training which was included in the PD workshop 2 (see Table 6.3). Moreover, I indicated very clearly the instructions of how they should conduct the session (for the instructions, see Appendix H). The first think-aloud session was carried out before the main activity of the PD; that is, before the revision of the rating criteria. This would prevent the PD from influencing how teachers rated the performances. The second session was carried out while implementing the PD workshop, and the final one after the last workshop. The excerpts of the think-aloud were used in the PD workshop as prompts for discussion (see also Table 6.3). The aim was to illustrate to the participants the differences among them in interpreting and applying the rating criteria as well as how each participant was diverse in terms of his or her own ways of rating students' performances.

5.2 Research Process

The data collection stretched over the period of a year and a half, which was carried out in three different phases: pilot study, main study, and follow-up study. Table 5.1 below illustrates the timeline of these phases, the principle objectives of each phase, and data collection methods employed. In this section, I describe the process of the main study. The pilot study is reported in Section 4.2, and the follow-up study will be reported in Section 6.3. In addition, I discuss the issues of reliability and validity of qualitative research, how I ensure the quality of my study, and how I took ethical issues in qualitative research into consideration.

Table 5.1: Data collection time frame (December 2006 – July 2008)

Pilot study (December 2006 – February 2007)

• Needs and problem analysis
  o Field observations
  o Interview 1
  o Interview 2
• Justifications and implications for the main study

Main study (October 2007 – February 2008)

• Implementing a PD (a series of 9 workshops)
• Examining the impact of the PD
  o Field observations
  o Interview 1; before participating in the PD workshop
  o Interview 2; during the PD workshop
  o Interview 3; after the final PD workshop
  o Focus group interview 1; integrated in first PD workshop
  o Focus group interview 2; integrated in last PD workshop

Follow-up study (June 2008 – July 2009)

• Confirming findings from the main study
  o Field observations
  o Interview 1; at beginning of the semester
  o Interview 2; after the first assessment task

However, before I started the main study, I applied the multi-facet Rasch measurement in examining the behaviours of teachers when rating students' performances, as a preparation for the study.

5.2.1 Preparing for the main study: investigating rater behaviours

The purpose of the preparation stage was to recruit the participants for the study. Because I wanted participation in the PD to be voluntarily, I first needed to illustrate the evidence to support the rationale of the PD and convince teachers of the value of getting involved in the extra work required. Because the Department did not keep students' performances or report any statistical data such as rater reliability in the performance-based assessment, teachers did not have an opportunity to learn about the reliability of their ratings. However, the results from the pilot study (see Sections 4.2.5.1 and 5.2.5.2) indicates low rater reliability as the teachers who participated in the study had different attitudes and beliefs toward rating criteria and their reported practices in their ratings. Therefore, I decided to use to the problem of low inter-rater reliability to make teachers aware of this problem and show the potential benefits of the project to the Department.

5.2.1.1 Data collection process

In this stage I started out by checking the inter-rater reliability of the scores of the students' written performances of Task 3, FE 1, offered in the previous semester. I used Task 3 because it was the final task of the course and some teachers did not return the tasks to the students. Task 3 of FE 1 requires students, in a group of three, to choose articles about problems and solutions from any available sources (e.g. the internet). For a written assignment for the task, students had to prepare a graphic organiser based on these articles for a written performance, and, for an oral performance, they had to give a presentation on these problems and solutions.

Since the Department did not keep students' performances or report any statistical data for the performance-based assessment (e.g. reliability), I had to ask around if any teachers kept students' FE 1, Task 3 written performances. A few teachers gave me their students' work of which I made copies. Then I invited a few

teachers to volunteer to rate these performances. Six teachers agreed to participate. These teachers were randomly selected. However, two teachers participated in the pilot study and two teachers would participate in the main study. I planned to have teachers rate 50 performances, but the teachers only agreed to rate 30 performances. Therefore, I randomly chose 30 performances from approximately 100 performances I had collected from teachers. I then made 6 copies of each performance so that each performance would be rated 6 times by 6 teachers.

In the morning of the agreed date, I explained to the participants in detail the purpose of this activity. After that, I gave them the rating scale (the same one they used when they rated in the previous semester) and explained that they had to follow the criteria. Then, I went through the rating scale with them. I also provided the participants with a grade record sheet. The participants decided that each of them would do the rating separately and return the score record sheet to me that evening.

In analysing the data, I decided to use multifaceted Rasch measurement because multifaceted Rasch measurement, one of several models developed within item response theory, can identify particular elements within a facet or aspect in performance-based assessment which is problematic, for example, a rater who is unsystematically inconsistent in his or her ratings. These facets include the ability of the candidate, the difficulty of the task, and characteristics of the rater. Multifaceted Rasch measurement has been used widely in the field of language testing and assessment (for examples of studies employing Rasch analysis, see Sections 2.4.1 and 2.4.2). For more detailed discussions on the implications of the multifaceted Rasch analysis in language testing and assessment, see McNamara (1996) (see also Myford & Wolfe 2003, 2004). It should be noted that though only 30 samples of performances and six teachers were used in this investigation, which did not represent the whole population of the students and teachers; the results were indicative of

teachers' inconsistent behaviours in making high-stake decisions on students' performances.

5.2.1.2 Results

When I received all the papers from the participants, I used the multifaceted Rasch analysis computer programme MINIFAC, a student evaluation version of FACETS (Linacre, 1989-2008), to analyse the data. Figure 5.1 below shows graphically the measures for students, raters, and traits from the rating scale.

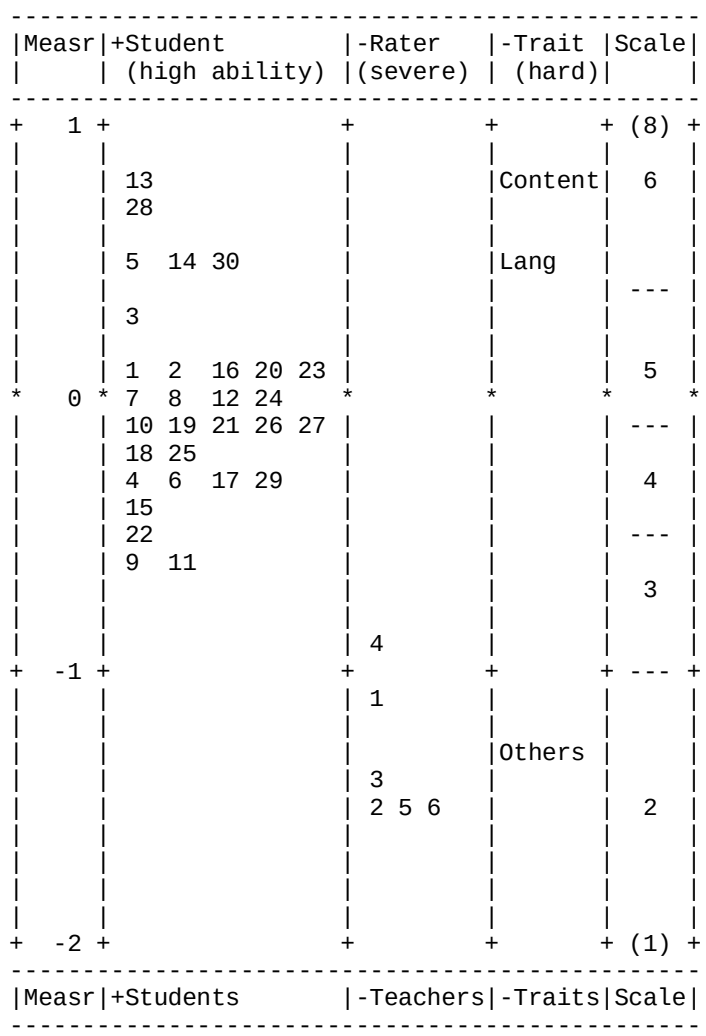


Figure 5.1: All-facet ruler summary

The figure is to be interpreted as follows. Students are ordered with the most able students at the top and the least able at the bottom. In terms of raters, the most severe rater is the uppermost rater in the figure. Likewise, the most difficult trait from the

rating scale is in the uppermost, and the least difficult trait is the bottom of the figure. As the figure indicates, raters disperse below the mean (0) of which 5 raters are with more than 1 logit measurement; that is, they are very lenient. For the rating scale, the trait Others is much below the mean and far less difficult than Content and Language. That is, raters are very generous in awarding high scores for the trait Others. On the other hand, the traits Content and Language disperse above the mean; that is, the three traits are different in their levels of difficulty.

Furthermore, the FACETS analysis provides estimates of, for example, examinee ability, rater harshness, and difficulty of the trait on a common log-linear metric. For the purpose of this study, only rater harshness/severity and trait difficult are reported. Studies employing the multi-faceted Rasch analysis are also discussed in Sections 2.4.1 and 2.4.2. The raters' measurement report is illustrated in Table 5.2 and the difficulty measure of the trait from the rating scale is shown in Table 5.3 below. The FACETS analysis provides a number of indications of the magnitude of the severity among raters.

Table 5.2: Raters' measurement report

Raters	Measure logit	Infit MnSq
R4	-0.88	1.76
R1	-1.11	0.75
R3	-1.36	0.85
R6	-1.46	0.70
R2	-1.51	0.75
R5	-1.54	0.75
RMSE:0.10 Adj. S.D. 0.22 Separation 2.15 Reliability:0.82		
Fixed (all same) chi-square: 38.5 d.f.: 5 Significance: .00		

The first indication of rater severity is the Separation Index. The Separation index is the ratio of the Adj. S.D. (corrected standard deviation) of the raters to the RMSE (root mean-square standard error). If the raters were equally severe, the Adj. S.D. should be equal to or smaller than RMSE. However, the rater Separation Index for the entire sample of raters is 2.15, indicating that the variance among the raters is about two times the error of estimates. Another indication is the Reliability statistic. This

Reliability statistic is not inter-rater reliability, but the degree to which the analysis reliably distinguishes between different levels of severity among different raters. Thus, low reliability is desirable, as ideally the different raters would be equally severe. However, in this case, the reliability is 0.82 indicating that the analysis is quite reliable in separating raters into different levels of severity. In summary, the raters are fairly different in their severity.

For the traits from the rating scale facet, similarly to the rater facet, the FACETS analysis also provides the same indications of the magnitude of the difficulty among different traits.

Table 5.3: Traits' measurement report

Trait	Measure logit	Infit MnSq
Content	0.81	0.86
Grammar	0.45	0.81
Others	-1.26	1.79
RMSE:0.08 Adj. S.D. 0.90 Separation 11.19 Reliability:0.99		
Fixed (all same) chi-square: 288.7 d.f.: 2 Significance: .00		

The trait Separation Index for the entire sample of traits is 11.19, indicating that the levels of difficulty among the three traits is about 11 times the error of estimates. For the Reliability statistic, the value of 0.99 signifies that the analysis is reliable in separating different levels of difficulty among the traits. In other words, the traits Content, Grammar and Others from the rating scale are very different in terms of their levels of difficulty.

5.2.1.3 Summary

The results from the multi-faceted Rasch measurement revealed that the raters were very different in their severity in rating students' performances, and the traits from the rating scale are significantly different in their levels of difficulty. Before implementing the PD workshop for the main study, I reported the results to the Department at a meeting in order to make them aware of the serious problems concerning the raters and the rating scale in the assessment of the course. In addition,

I reported the results to the participating teachers in the introduction session of the PD workshop. These teachers later agreed that the rating criteria were one of the major problems of the assessment process of the foundation courses. Therefore, they decided to accept a revision of the criteria as the main focus of the PD workshop.

5.2.2 Main study

Drawing from the findings from the pilot study and literature review, the aims of the main study were to implement a PD in language assessment for teachers and examine the impact of the PD on the teachers who participated in the programme. The main study began in October 2007, which was the beginning of Semester 2. In this semester, FE 2 was offered (FE 1 was offered in Semester 1). Therefore, in the PD, the course investigated was FE 2. For the assessment, due to the nature of oral assessment which is a live presentation, in the present study I did not have adequate resources and time to collect sample performances. Thus, the written performance of Task 1 was used in the PD. Moreover, based on the results from the preparation stage discussed above and the participants' agreement, the PD activities were mainly focused on revising the criteria for this task.

5.2.2.1 Purposes of the study

To gain an in-depth understanding of the impact of PD on teachers, it is crucial to:

1 Understand teachers' beliefs, attitudes and knowledge in language assessment
2 Understand the relationships between the above constructs and what teachers do
3 Understand how the PD programme affects teachers' beliefs, attitudes and knowledge and their practices in language assessment
4 Discover whether teacher's beliefs, attitude, knowledge, and practices in assessment change as the result of a PD programme

5.2.2.2 Research questions

The primary research question addressed in this study is: How does an in-service PD programme in language assessment affect classroom EFL teachers' beliefs, attitudes, knowledge, and practices in relation to assessment?

This question will be addressed through consideration of a number of more specific, subsidiary questions:

Before implementing the professional development programme:

1 What are the teachers' beliefs, attitudes and knowledge about language assessment and the assessments being used?

2 In what way do their beliefs, attitude and knowledge influence what they do?

After implementing the professional development programme:

1. Have the teachers' beliefs, attitudes, and knowledge in language assessment changed after the PD programme? If yes, in what way? If no, why?
2. Have the teachers' assessment practices changed after the PD programme? If yes, in what way? If no, why?

5.2.2.3 Data collection processes

The process in collecting the data for the main study can be divided into three stages: before implementing the PD workshop, while implementing the PD workshop, and after implementing the PD workshop.

Stage 1: Before implementing the professional development workshop

A semi-structured interview method was employed in order to find out about the participating teachers' knowledge about language assessment, their beliefs about it and attitudes towards it. The interviews also allowed the teachers to describe how they do the assessment. The first individual interviews were conducted during the

week before the PD workshop 1. Each teacher spent about 20-30 minutes answering the questions. The interview was focused on the teachers' views towards language assessment in general, the assessments used in the course, as well as how they did the rating in the previous semester. (For guiding questions, see Appendix K)

In addition, a focus group interview was carried out in PD workshop 2. The participants were asked to critique the assessments being used with reference to the course's objectives and syllabus. The participants were also asked to do a think-aloud for each of the written assessment tasks done in classroom. There are 3 written assessment tasks in this course. The purpose of using think-aloud was to study the participants' underlying psychological processes as well as their practices while they did the rating. The first think-aloud was done prior to the main activity of the PD (see Section 5.1.2.4). The participants were trained on the think-aloud protocol before they actually did the think-aloud (for the think-aloud instructions, see Appendix H).

Stage 2: While implementing the professional development workshop

Semi-structured interviews were employed in this stage in order to understand what teachers believe, along with their attitude and knowledge about language assessment (for guiding questions, see Appendix K). These interviews also focused on what the participants had learned from the workshops as well as their views towards the workshops. Second, think-aloud protocols were also carried out at the same week of the interviews.

Stage 3: After the professional development workshop

The last PD workshop was conducted as a focus group in which teachers were encouraged to share their views toward the assessment used in the course, which included the assessment tasks and their criteria and scale. In the focus group, stimuli, which consisted of materials used and produced in the PD, were used to recall the activities carried out in the PD as well as the course's assessment. This focus group

also provided teachers an opportunity to evaluate the PD and prepare for the next phase of the research.

After the last PD workshop, an individual semi-structured interview with stimulated verbal recall method (the same set of stimuli used in the focus group was applied) was employed in order to find out about the participating teachers' knowledge concerning language assessment, their beliefs about it and attitudes towards it. The interviews allowed the teachers to report how they do the assessment. In addition, these interviews illustrated the changes of the teachers. (For guiding questions, see Appendix K). Furthermore, the third think-aloud protocols were conducted after the final PD workshop.

5.2.2.4 Participant profiles

The participants in the main study and the follow-up study were the same teachers, which included 5 teachers who were teaching FE courses at the time of the study. Though they were selected with opportunity and convenience taken into account, I was successful in recruiting participants from different background, such as gender, education, and teaching experience.

Teacher 1: Catbandit

Catbandit is 29 years old. He started his teaching career at the English department Chiang Mai University where he received his BA in English in 1999. After having taught EFL for 3 years, Catbandit pursued an MA in Linguistics at Chulalongkorn University, Bangkok, Thailand which he completed in 2005. Upon his completion, he resumed his teaching position at Chiang Mai University where he has served as a coordinator of FE 1.

Teacher 2: Papone

Papone is 34 years old. He did his BA in English at Payap University, a private university in Chiang Mai, Thailand. Papone started his teaching career at a language school in Chiang Mai where he had taught for a year before he went back to Payap University for his MA in TEFL. Papone then moved to Bangkok to start teaching EFL at Srinakharinwirot University, where he taught for 2 years. After that he accepted a teaching position at Chiang Mai University. Papone served as a coordinator of the previous fundamental English 1, and has been a coordinator of FE 2 since it was first implemented. In addition, he was one of the material developers of this course.

Teacher 3: Songsri

Songsri is 49 years old. She is the most experienced teachers among the participants. She received her BEd in English from Chulalongkorn University, Thailand in 1981 when she started teaching EFL at King Mongkut's University of Technology North Bangkok. After having taught there for 3 years, Songsri went to King Mongkut's University of Technology Thonburi, Thailand to pursue an MA in Applied Linguistics (English for Science and Technology). Upon her completion in 1993, Songsri continued her teaching career at Chiang Mai University where she has been involved in English for Science and Technology courses. She wrote about 75% of the English for Science and Technology 1 textbook, which was offered prior to the new foundation courses. Songsri was also one of FE 4 (English for Science and Technology) material developers. She had considerable experiences in test developing and writing.

Teacher 4: Tanya

Tanya is 32 years old. She finished her BA (German) at Chiang Mai University, Thailand in 1998 when she went to Bangkok to work as a secretary. After 2 years,

Tanya resigned to work as an international relation officer for a national television company. In 2003, Tanya enrolled in an MA programme in TEFL at Thammasart University, part-time. Upon completion in 2005, she went to the University of Nottingham, UK, to do another MA in Applied Linguistics. Tanya started her teaching career as a part-time instructor at Chiang Mai University. After having taught for one semester, she was offered a full-time position. Therefore, Tanya had only 5 months teaching experience prior to this study. She started her full-time position after 2 workshops.

Teacher 5: Wanwisa

Wanwisa is 34 years old. She received her BA in English at Konkaen University, Thailand in 1996 after which she started her teaching career at Chiang Mai University. In 1998, Wanwisa went to Silapakorn University, Bangkok, to further her education in Med (English), but she resigned after one semester because she found that the programme was not what she wanted to do. Wanwisa then resumed her teaching at Chiang Mai University. In year 2000, Wanwisa went to Thammasart University, Bangkok, to do an MA in English (Literature). After one semester, however, she was awarded the grant, AusAid, offered by the Australian government, to continue her education in Australia. She earned an MA in Applied Linguistics from University of Western Australia in 2002 and has been back at Chiang Mai University since then. Wanwisa was one of the material developers of FE 4 (English for Humanities and Social Sciences).

5.2.3 Issues of Validity and Reliability of the Qualitative analysis

In this section I take into account different criteria proposed by scholars in order to ensure the quality of qualitative research. I do not present the debate on the validity and reliability of qualitative research from the ‘paradigm war’ point of view but

practical steps or guidelines to carry out the research. Dörnyei (2007, pp. 59 - 62) summarises several strategies that have been proposed to ensure the quality of qualitative research, consisting of:

- *Building up an image of researcher integrity* through audit trails, contextualisation and thick description, identifying potential researcher bias or examining outliers, extreme or negative cases and alternatives explanations.
- *Validity/reliability checks* by incorporating respondent feedback and member and/or peer checking into research designs.
- *Research design-based strategies* which consist of method and data triangulation, prolonged engagement and persistent observation and longitudinal research designs.

Based on these strategies, I adopted different steps to demonstrate the reliability of the analytical process and the validity of the claims made in this thesis, including:

1 providing an 'audit trail', which is 'created by documentation of the research process and by provision of sufficient evidence to understand how the researcher reached the conclusion of the study' (Morrison & Hamp-Lyons, 2007), in Section 5.3.2.2;

2 member checking, that is sending my overview of the data (Chapter 6) to the participants and asking them to critically analyse and comment on the data (for email exchange, see Appendix L);

3 offering a detailed description of research methodology, in the previous chapter;

4 providing the detailed description of my roles of the researcher, in the following section; and

5 collecting the data during a series of points in time; in other words, being a longitudinal research.

5.2.4 Roles of the researcher

It is crucial to note here my role as a researcher in the data collection process because it directly affects the data as well as the data interpretation. First of all, as I pointed out in the introduction of the thesis, I have been an instructor in the Department since 2000. I was recognised by the research participants and teachers in the Department as a member of the staff who was on study leave and was conducting a PhD research project in language testing and assessment. In addition, they were aware that my research project aimed to improve the quality of assessment in the Department. Therefore, the teachers were very keen and willing to collaborate. For instance, I was allowed to attend any meetings I requested and use the rooms at the Department for the PD workshop. Furthermore, my presence in the Department, for instance, in the teacher's Common Room was perceived as an ordinary circumstance. In other words, I was accepted as part of the community.

It should be noted that I was one of the six teachers who developed the materials and assessment for FE 1 and FE 2. During the data collection, I was recognised as one of the material development team. The teachers in the interviews often recalled this fact and assumed that I already understood what they were talking about. In order to make sure that the data I collected for the further interpretation and discussion of the data is not biased, I had to tell these teachers that they had to speak to me as if I did not know anything about the course. For example, in an interview when the interviewee said "*You know what I mean because you were there*", I had to redirect "*What do you mean*?" My roles in the PD workshop are described in Section 6.1.2.

5.2.5 Ethical Issues

In this session, I review ethical issues proposed by researchers and how I took these issues into consideration when I conducted the study. Drawing from Lipson (1994), Creswell (2007) categorises ethical issues into 'informed consent procedures;

deception or covert activities; confidentiality toward participants, sponsors, and colleagues; benefits of research to participants over risks; and participant requests that go beyond social norms' (p. 141). Moreover, the American Anthropological Association has specified the following standards: a researcher protects the anonymity of the informants, a researcher develops case studies of individuals that represent a composite picture rather than an individual picture, and a researcher conveys to participants that they are participating in a study, explains the purpose of the study, and does not engage in deception about the nature of the study (Creswell, ibid., p. 142). Furthermore, Cohen, Manion and Morrison (2007) categorise the ethical principles for educational research into two categories: responsibility to research and responsibility to participants and audiences. These principles should be agreed upon 'before' the research commences (for more detail, see Cohen et al., ibid., p. 77).

The main study and follow-up study included 5 participants who were full-time Thai teachers and teach FE courses (except 1 participant who was a part-time teacher at the beginning of the study but became full-time later on). When I started the present study at the Department, I started asking teachers, whom I considered as friends, if they would be interested to participating in my study. I told them that the study would require them to participate in approximately 9 workshops and 3 interview sections. Though the participants were selected with opportunity and convenience taken into account, I managed to recruit participants from different backgrounds (for detail of participant profiles, see section 5.2.2.4).

Furthermore, when Professor Liz Hamp-Lyons, my doctoral supervisor, visited the Department at the beginning of the main study (26 October 2007) and gave a one day workshop on performance-based language assessment, she suggested to the teachers in the workshop that my research would be beneficial for the Department. After the workshop, five teachers agreed to fully participate. Before the

first PD workshop, I explained to these teachers in detail the procedures, guidelines, and confidentiality of the data (see Appendix I). I also asked them to sign and keep a copy of the Consent Form (see Appendix J). In terms of anonymity of the participants, I asked them to choose the names they would like me to refer to them in the thesis. In addition, after I had written the overview of the data, I sent the participants an email asking them to validate the content of the data. I also asked them if they wanted to take out any parts and whether they wanted to change their identities. When I received their replies I did according to their requests. The same procedures were also used in the pilot study.

5.3 Data Analysis

The data for the analysis includes the three interviews with each participant from the main study, two focus group interviews, and two interviews with each participant in the follow-up study. The analysis of the data was guided by the Grounded Theory, which is presented below. The process of the analysis of the data is described in Section 5.3.2 below.

5.3.1 Grounded Theory for Data Analysis and Interpretation

As described in Section 5.3.2.2, the present study employed Grounded Theory (GT) as a tool for data analysis and interpretation, and in this section I describe briefly how GT can be used for this purpose to achieve theory building. Though, as pointed out by Dey (2004) there are many versions of GT, for example, Glaser (1987), Strauss (1987), Charmaz (1990), and Strauss and Corbin (1990); this study follows Strauss and Corbin's version. The explanation of this section is exclusively drawn from Corbin and Strauss's (2008) most recent work *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory* (3rd Edition). In doing this, I put together the ideas and concepts from different components and chapters of

the books to present how I employed GT in the process of data analysis. In the following discussions I explain the techniques of employing GT as strategies for qualitative data analysis, which comprise of coding (open coding and axial coding), integrating categories and theory building, and memoing.

5.3.1.1 Coding

Corbin and Strauss (2008) describe that in GT, doing analysis involves coding which is the process of generating, developing, and verifying concepts. They emphasise that coding is more than just a paraphrasing, noting concepts in the margins of the field notes or making a list of codes as in a computer programme. Coding, in other words, is the process of interpreting the data. There are two types coding: open coding and axial coding.

Open coding: analyzing data for concepts

According to Corbin and Strauss (ibid., p. 160), doing analysis starts with 'open coding'. Open coding requires a brainstorming approach to analysis in order to open up the data to all potentials and possibilities contained within them. In this process, the researchers, after having considered all possible meanings, put interpretive conceptual labels on the data. Corbin and Strauss emphasise that these concepts represent the researchers' impressionistic understanding of what is being described by the participants. In addition, they describe that concepts can range from lower-level concepts to higher-level concepts. Higher-level concepts are called categories/theme and categories tell us what a group of lower-level concepts are pointing to or are indicating.

Furthermore, Corbin and Strauss (ibid., p. 160) also provide the steps in constructing concepts including:

1 break the data into manageable pieces,

2 take those pieces of data and explore them for ideas contained within (interpreting those data,) and

3 give those ideas conceptual names that stand for and represent the ideas contained in the data.

Axial coding: elaborating the analysis

Though open coding and axial coding are treated as if they occurred separately, Corbin and Strauss (ibid., pp. 198 - 199) point out that the distinctions made between the two types of coding are artificial and for explanatory purposes only. They also stress that whereas open coding is breaking data apart and delineating concepts to stand for blocks of raw data, axial coding is the act of relating concepts/categories to each other. They explain that in the process of open coding, while the researchers break data apart and identify concepts to stand for the data, in their minds, they automatically put the data back together and make connections by creating the explanatory descriptors – doing axial coding. In other words, open coding and axial coding occur concurrently. In linking the categories and making connections among them, the researchers also elaborate on them. Linking could occur from a lower-level to a higher-level, similar to linking blocks to build a pyramid. Corbin and Strauss stress that elaborating on the analysis is the process in which the researcher explains this pyramid by explaining these blocks and how they are arranged.

5.3.1.2 Integrating categories and theory building

According to Corbin and Strauss (2008), the first step in integration is deciding upon a ‘central’ or ‘core’ category, which represents the main theme of the research. It is the concept that all other concepts are related to. In other words, it is the category that appears to have the greatest explanatory relevance and highest potential for linking all of the other categories together. The following step is refining the theory. Corbin and Strauss explain that theory building is a process of going from raw data to making

statements of relationship about those concepts and linking them all together into a theoretical whole.

5.3.1.3 Memoing

Corbin & Strauss (ibid., p. 117) stress that memos are a specialised type of written records – those that contain the products of the analyses. Writing memos should begin with the first analytic session and continue throughout the analytic process. It is part of the analysis, part of doing qualitative research because they move the analysis forward. Memos are rudimentary representations of thought and grow in complexity, density, clarity, and accuracy as the research progresses. For a sample memo, see Figure 5.4 below.

5.3.2 Analysing the data

Following Corbin and Strauss's (2008) analytical guidelines in analysing qualitative data discussed above, in this section I describe the steps I took in analysing and interpreting the data.

5.3.2.1 Data storage and transcription

All the interviews (in Thai) were digitally recorded and stored electronically as sound files under the file name which included the date (for easy identification purposes) in separate document folders allocated to individual participants (for illustration, see Figure 5.2). Interviews were transcribed verbatim and their summaries were typed up as word documents. A summary of each PD workshop and observations from the meetings were word processed.

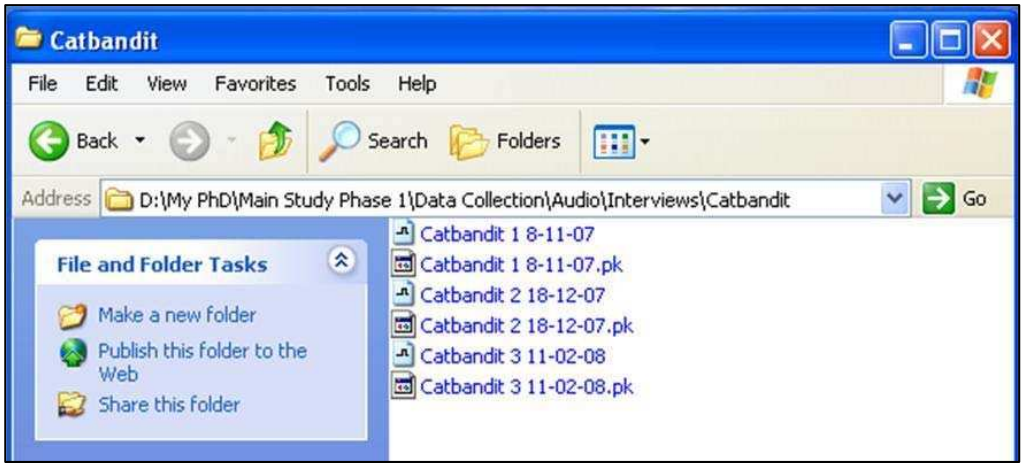


Figure 5.2: A computer screen shot displaying the storage of sound files of an individual participant

5.3.2.2 Coding, integrating categories, and theory building

The process of initial analysis involved listening to the interviews of each participant a few times to acquire a fresh the memory of the interviews. After listening to the interviews, I read the transcripts and typed up a summary of each of the interviews. The listening and summarising helped me to have a deeper understanding of the interviews. After that, I read the transcripts again and paid particular interest in coming up with codes and possible categories. When I found an incident which was interesting, or related to teaching and assessment, I underlined that incident and tried to understand what it meant to come up with a code, therefore, doing open coding. Then, I would write down the code on the right margin of the transcript along with the summarised ideas for that particular code. Figure 5.3 below illustrates the codes on the transcript.

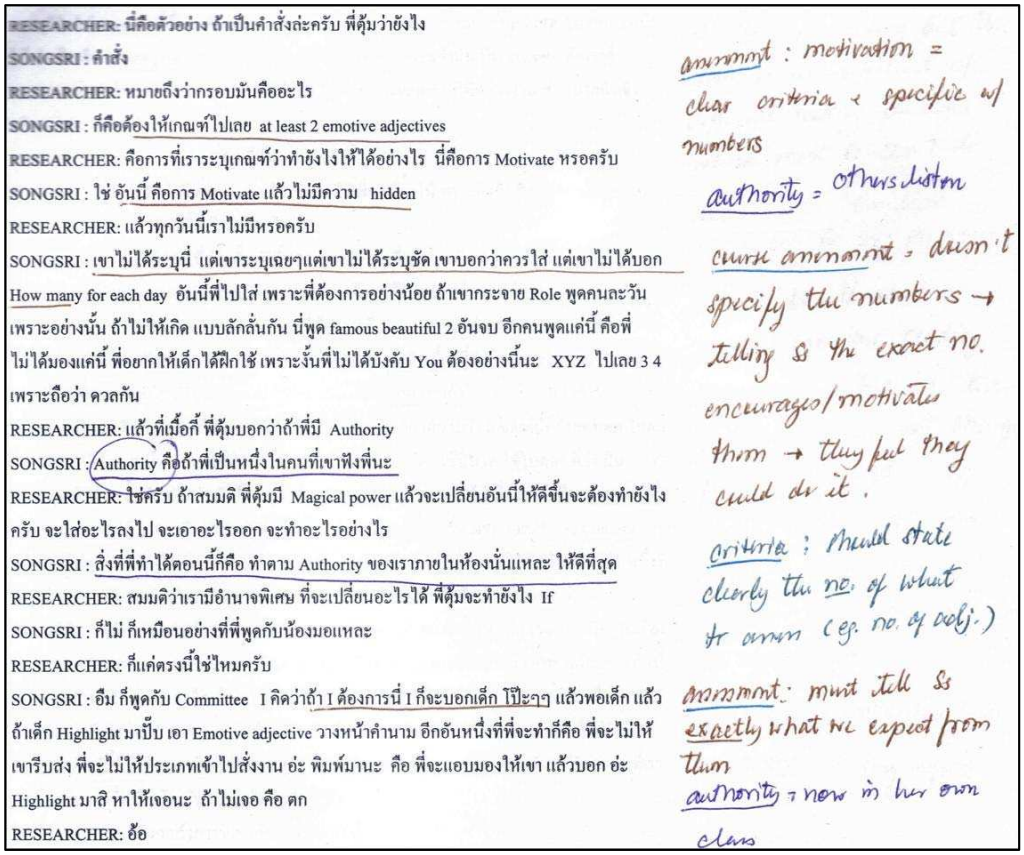


Figure 5.3: Sample of open coding on the transcript

When I finished coding one participant on the transcript, I wrote memos of that participant as a word document. In the memos (for a sample memo, see Figure 5.4), following Corbin and Strauss (2008, pp. 117 - 118, as described above), I wrote the memo number on the top, and I also noted on the left margin of the transcript the same number so that I could refer this memo back to the transcript (when I need to in the future). Under the memo number, I wrote the assigned code with the date coded. Then, I paraphrased the incident under the code. When I found that I had some comments about any paraphrased incidents, I also noted down my comments next to those incidents. While coding, when I saw some emerged categories, I would assign categories for the codes, or doing axial coding. Moreover, I made annotations (for a sample annotation, see Figure 5.5 below) when I felt there were interesting issues or themes emerging from the memos.

Memo 1

1 September 2008
Belief in assessment
Traditional exam VS Performance assessment
Tanya thinks that traditional exam assesses student's competence, including memorisation and grammar. It is 'standard' and easy to mark. On the other hand, performance assessment assesses student's performances. Thus, the current assessment for the foundation courses assesses both competence and performance because the courses consist of final exam and performance assessment. Tanya also adds that some students are good at competence whereas some are good at performance. *The question is whether the exams they use in the departments are 'standard' since they do not have any measure in standardising the exams. Perhaps, what Tanya means by 'standard' is that there is a standard marking, that is, objective marking.*

Figure 5.4: Sample memo

Annotation

1 September 2008
New teacher
Enthusiastic in learning
It is important to note that Tanya is a new teacher. She has only been teaching at the department for 1 semester. This is also her first year of teaching career. During the time of the interview, she was holding part-time position. However, she has just gone through the assessment process of being a full-time, in which she would know the result that she passed in a few weeks' times. As being a part-timer, Tanya is very enthusiastic in learning as reflected by the fact that she wants to become a full-time and has decided to participate in this PD. Generally speaking, part-time teachers do not engage in academic activities in the department.

Figure 5.5: Sample annotation

Moreover, it is worth noting that the transcripts I worked on are in Thai, since the interviews were done in Thai. I did not translate the transcripts before the coding process because I believe that working with authentic texts would give me richer information. However, I paraphrased the coded incidents into English because these excerpts might be included in my thesis.

After I finished coding the interviews of two participants, I wanted to have a better understanding of the emerged codes and categories. Therefore, I printed out all the memos. Resorting to the traditional paper methods, I categorised them by putting

the memos with similar theme in the same piles, thus, the initial phase of integrating categories. What I discovered was that I needed to make changes to the codes and categories I had assigned because these codes and categories were created when I started coding the first participants four months earlier. Consequently, I made changes of the titles of the codes of the first two participants I had previously coded. When I finished coding all participants, I imported the codes into NVivo 7 software (QSR International, 2006). While importing the codes from the word documents into NVivo, I found that my consideration of the data had been more defined, as I was more familiar with the data. Therefore, I changed the wording of the codes and the categories I previously made, as well as created hierarchical relationships between the codes and categories. In other words, it was an integrating categorising process. More importantly, I realised that I had to use a different approach in analysing the data, that is to re-code the interview transcriptions.

After I had finished the importing, I went on further to do more investigation of the data in an in-depth analytical manner by comparing and contrasting the structures of the codes. Using NVivo was a great advantage because the software could illustrate the tree nodes (*node* is a term for *codes* used in NVivo) of the coding scheme. The figure below represents the coding structure of the first analysis. Moreover, I used the 'models' facility of NVivo to have a visual representation of the categories and codes, which would later help with the coding tree structure outputs, refined the categories and codes. Figure 6.4 below is an example of a model created by NVivo as a map.

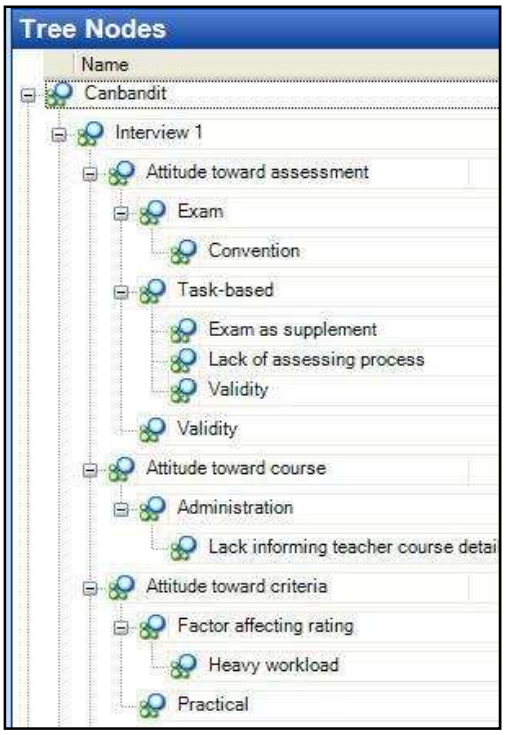


Figure 5.6: Sample of NVivo output of a coding tree structure

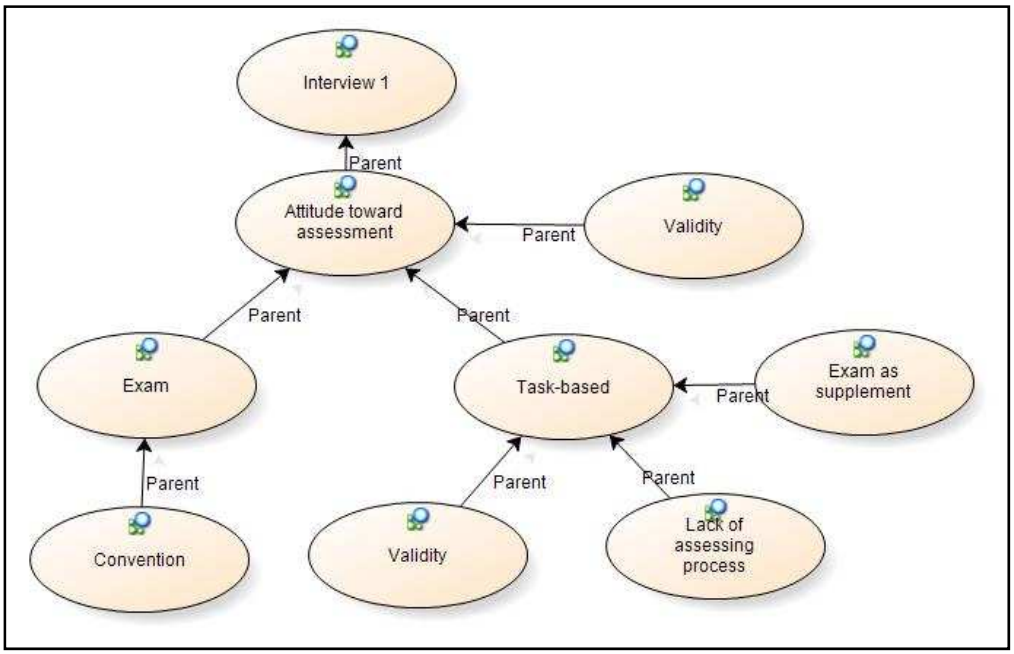


Figure 5.7: Sample of NVivo output model of categories and codes as a map

After I had studied both tree nodes and map representations of the coding structures, I edited some codes and categories by making changes to the titles as well as moving some codes to appropriate categories. Then I re-coded the interview

transcripts using a theme approach. For this coding, I started off with a theme/ category derived from the previous analysis in mind, and then went through the freshly printed transcripts and coded all participants for that particular theme; in contrast to the previous coding in which I did not have any theme in mind. When I finished coding the data on the transcripts, I transferred the codes into NVivo. Then I repeated the same analytic procedures in comparing and contrasting the codes and categories. After having done rigorous analysis, thus, another further step of integrating categories, I came up with a new set of codes and categories. Table 5.4 below shows the comparison of the codes and categories from the first and second coding. I relied on these codes and categories, with the aids of NVivo, in writing up the overview of the data. In addition, the interpretation and discussion of the findings, thus, are based on the second set of codes. The overview and the interpretation of the data will be presented in Chapter 6. The steps of theory building will be illustrated in the introduction of the discussion chapter, Chapter 7.

Table 5. 4: Comparing codes from first and second coding

First Coding	Second Coding
Interview 1 *Attitude toward assessment* • Exam o Convention *Task-based* o Exam as supplement o Lack of assessing process o Validity *Attitude toward course* o Administration ▪ Lack informing teacher course detail *Attitude toward criteria* *Factor affecting rating* o Heavy workload o Practical *Reported practice* o Impressionistic rating o Learning from experience o Using impression *Teacher's responsibility* o Awareness on syllabus & assessment o Submitting scores timely	**Interview 1** *Attitude toward assessment* • Course assessment o Exam ▪ Necessary ▪ Assessable ▪ Conventions ▪ Supplement to task-based o Management ▪ Punctuality o Objectives o Task performance-based ▪ Lack assessing process ▪ Strategies ▪ Validity • Questionable • Assessment in general o Course objectives ▪ How ▪ Validity • How • What • Weighting o Workload ▪ Rating • Decrease quality *Reported practice* • Criteria o Follow o Not follow ▪ Experience ▪ Impression

5.4 Conclusion

This study is a longitudinal qualitative research which integrated grounded theory, ethnography and case study approaches. I used a grounded theory approach because this approach allows me to derive the data from the views of each participant. It is an

ethnographic study as I was in the research setting as an observer to explore the culture of the Department and the participants' shared experiences, attitudes, knowledge, beliefs. Moreover, I adopted a case study approach as it could provide a thick description of the situation of each participant. In collecting the data, I employed interviews, focus group and think-aloud methods as my methodological tools.

In this chapter, I have also described in detail the main study, including: purposes, research questions, data collection process, and participants of each study. In addition, I include a brief summary of the Grounded Theory components employed in the data analysis and interpretation. I also provided a detailed account of the procedures involved in the analysis of the data to provide a transparent picture of the procedure and therefore to demonstrate reliability and validity of the analysis and interpretation. The final section deals with ethical issues. The following chapters are the overview of the data (Chapter 6) and the discussion of the findings (Chapter 7).

6 Professional Development, and Teacher's Thinking and Reported Practice in Assessment: Data Overview

In this chapter, I present the data overview from the main study and the follow-up study. In the first part of the chapter, I offer the brief information regarding the professional development (PD) workshops; including the purposes, the structure and the activities carried out in each workshop. The second part of this chapter is devoted to the exploration of the data collected from the interviews with the five teachers who participated in the PD workshop. The last part of the chapter reports the findings obtained from the follow-up study. The discussion of the data will be explored in the following chapter (Chapter 7).

6.1 The Professional Development Workshops

The findings from the pilot study (see Sections 4.2.5.1 and 4.2.5.2) and the investigation of rater behaviours in the preparation stage of the main study (see Section 5.2.1.2) indicate that, the problems in assessment in the Department were mainly caused by the teachers' various attitudes towards the rating criteria and their inconsistency in the ratings. In addition, the findings from the pilot study reveal that another cause of the problems pertaining to assessment was the lack of teachers' adequate understanding of performance-based language assessment (see Section 4.2.5.3). From reviewing literature in general, and language education, as well as language testing and assessment, a PD programme has been proposed as one of the solutions to this type of problems (e.g. Brindley, 2001; Hamp-Lyons, 2007b; Malone, 2008; Pillay, 2002; Prapphal, 2008). Therefore, providing teachers with PD, and the investigation of the development of these teachers, became the focus of the main study of this research project. The PD programme, consisting of nine meetings with

five teachers, was carried out at the English department over a period of four months from November 2006 to February 2007. The data collections were also conducted concurrently (see Table 6.2 below).

6.1.1 Purposes

The primary objective of the PD was to provide theoretical and practical aspects of performance-based language assessment to teachers. For the theoretical aspect, the focus was on the rating process, with the emphasis on the rater and rating criteria. For the practical part, in response to the problems concerning rating criteria, the PD aimed to offer the teachers hands-on experience in revising and developing rating criteria for a performance task.

6.1.2 Structure

To achieve the above objectives, the PD was carried out as a series of nine workshops which focused on both theoretical and practical aspects of performance-based language assessment. Each meeting lasted one to two hours, depending on the availability of the participants. The teachers who participated in the workshops took active roles in sharing opinions and making suggestions on the assessment brought into the discussions. They also took part in the debates when there were disagreements. My main roles in the PD workshop were facilitator and a discussion leader, apart from giving inputs in fundamental principles in performance-based assessment. I also prepared all of the materials for each workshop, as well as compiled and summarised the discussion from the previous workshop. For example, I listed the descriptors the participants suggested, and, grouped and inputted them into a rating scale.

While the main activity participants concentrated on was revising the rating criteria, providing fundamental principles of performance-based assessment focusing on rater and rating criteria was also integrated into the workshops. For instance, in

Workshop 4 (see Table 6.1 below) I introduced the concept of rater reliability to the participants. I achieved this by using excerpts from their think-aloud protocols to illustrate the differences of how the individual raters interpreted and applied the rating criteria differently. The concepts of 'inter' and 'intra' rater reliability were discussed. At the same time, an additional aim of this activity was to analyse the constructs each teacher considered while rating the sample performances (from the sample protocols). These constructs were complied as possible descriptors of the revised criteria. Table 6.1 (shown below) summarises the objectives, materials used, and the activities carried out in each of the PD workshop.

6.1.3 Workshop activities

In the workshops, the participants agreed that the rating criteria needed to be revised; therefore, the main activity of the workshop was to revise the criteria for a writing task of Task 1 FE 2. This task was used because FE 2 was the official course offered in that semester. In addition, because of time limitations, only one set of criteria was used. Furthermore, in the workshops I provided relevant principles in performance-based assessment when appropriate. In Table 6.1 below, I provide the primary objectives of each workshop and the material used. I also explain my role in each workshop, as well as the activities the participants undertook in each workshop.

Table 6.1: Summary of professional development workshops

Workshop 1(16-11-07)			
Objectives	*Materials*	*Activities*	
		Researcher	*Participants*

• Introduce the PD to the participants
  o Objectives
  o Procedures
	• Findings from the pilot study (cf. Section 4.2.4)
	• Results from investigating rater's behaviours (cf. Section 5.2.1.2)
	• Outlines and timetable of the PD workshop (see Appendix P)
		• Reported findings from the pilot study and the investigation of rater's behaviours
		• Outlined the research project
		• Introduced think-aloud
			• Asked questions about the project

Workshop 2 (23-11-07)			
Objectives	*Materials*	*Activities*	
		Researcher	*Participants*
• Evaluate the assessment tasks • Identify problems with the assessment • Practice think-aloud	• Assessment tasks (see Appendix E) • Rating criteria (see Appendix E) • Think-aloud instructions (see Appendix H)	• Asked teachers to comment on the assessment (using the assessment tasks and the rating scales from the course) • Invited the teachers to share their rating experiences	• Shared their opinions on assessment tasks and the rating criteria • Shared their rating experiences • Practiced think-aloud
Note: To facilitate the participants, I compiled the objectives of the course, objectives of each class period, as well as the assessment tasks and their rating criteria from the course syllabus (see Appendix O).			

Workshop 3 (30-11-07)			
Objectives	*Materials*	*Activities*	
		Researcher	*Participants*
• Further evaluate the assessment tasks • Introduce basic concepts in performance-based language assessment	• Course's syllabus and related course materials (see Appendix O) • Glossary of terminology relating to performance-based assessment (see Appendix T)	• Presented the course's syllabus and materials relating the assessment • Asked the participants to study and critique the assessment tasks, the criteria, and rating scales • Explained the definitions of important terminology in performance-based language assessment	• Shared their views toward the assessment and the course's syllabus • Identified problems with the objectives of the course and the assessment tasks • Discussed different aspects of the assessment and the criteria • Shared rating experiences • Concluded to revise the rating criteria for Task 1 because there were problems with the descriptors • Studied the terminology
Workshop 4 (14-12-07)			
Objectives	*Materials*	*Activities*	
		Researcher	*Participants*
• Introduce factors affecting rating: raters • Evaluate the existing rating criteria	• Excerpts from think-aloud protocols (see Appendix N) • Glossary of terminology relating to performance-based assessment • Rating criteria (Task 1, written task)	• Introduced the concepts of rater reliability • Presented the think-aloud protocol excerpts • Distributed comments on the criteria made in the last workshop	• Commented on the protocols • Studied previous workshop's comments • Compared the existing descriptors with the course syllabus • Suggested possible descriptors from their rating experiences

Note: After the workshop, I interviewed the participants (the second round of individual interviews) and found that they wanted to include a few more participants in the criteria revision process**.** Therefore, I invited 2 teachers to participate in the following workshop.

Workshop 5 (21-12-07)			
Objectives	*Materials*	*Activities*	
		Researcher	*Participants*
• Introduce factors affecting rating: rating scales • Evaluate the existing rating criteria by comparing them against the samples of performances	• Glossary of terminology relating to performance-based assessment • The rating criteria • List of descriptors (derived from previous workshop) (See Appendix R) • Samples of performances (see Appendix S)	• Introduce the concepts of analytic and holistic scales • Presented the complied list of categories and descriptors derived from the previous workshop • Invited the participants to study the sample performances and ask them to describe the performances of different levels	• Studied on the compiled list of descriptors and compared them with the sample performances • Shared their rating experiences • Commented on the descriptors
Note: Two other teachers participated in this workshop			
Workshop 6 (04-01-08)			
Objectives	*Materials*	*Activities*	
		Researcher	*Participants*
• Further explore components of rating scales • Revise the criteria	• The rating criteria • Samples of assessment criteria (see Appendix U) • List of descriptors and categories (derived from previous workshop) (See Appendix R)	• Studied samples of assessment criteria • Presented the list of descriptors and categories • Invited the participants to comment on the descriptors	• Commented and revised the descriptors based on their rating experiences • Rearranged/regrouped the descriptors

Workshop 7 (16-01-08)			
Objectives	*Materials*	*Activities*	
		Researcher	*Participants*
• Introduce steps in developing rating criteria • Further revise the criteria	• Steps in developing rating criteria (Appendix A) • Feedback from teachers on the revised criteria • The revised criteria(See Appendix R)	• Presented the rating scale derived from the previous workshop • Showed the feedback from teachers on the criteria	• Revised the scale based on the course's syllabus and their rating experiences • Studied the feedback and revised the criteria if they agreed
Workshop 8 (01-02-08)			
Objectives	*Materials*	*Activities*	
		Researcher	*Participants*
• Further revise the criteria	• The revised criteria(See Appendix R) • Samples of performances (see Appendix S)	• Led the revision • Finalised the revision • Concluded the process of designing rating criteria and what we had done	• Tried using the criteria in rating performance samples • Shared their opinions about using the criteria • Revised the criteria
Workshop 9 (09-02-08)			
Objectives	*Materials*	*Activities*	
		Researcher	*Participants*
• Summarise the principles in performance-based assessment • Conclude the workshop • Reflect on the workshops	• Slides from PowerPoint Presentation (see Appendix Q) • Materials used in the workshops	• Summarised the principles discussed in the workshops • Asked teachers to share their opinions on the assessment • Invited the teachers to share their views toward the PD	• Shared their opinions toward the assessment and the PD

As far as the data collection for the present research project is concerned, three individual interviews and two focus group interviews were carried out while the workshop was being conducted. Table 6.2 below illustrates the time frame of the PD workshop and the interviews (cf. Table 6.1 above).

Table 6.2: Data collection and the PD workshop time frame

PD workshop			**Data collection**	
No	Date	Main objectives	*Interviews*	*Date*
			Individuals round 1	*6-15/11/07*
1	16/11/07	Introduction		
2	23/11/07	Evaluating the assessment tasks	*Focus group 1*	*23/11/07*
3	30/11/07	Evaluating the rating criteria		
4	14/12/07	Understanding performance-based assessment & revise the rating criteria	*Individuals round 2*	*17-18/12/07*
5	21/12/07			
6	04/01/08			
7	16/01/08			
8	01/02/08			
9	09/02/08	Conclusion	*Focus group 2*	*09/02/08*
			Individuals round 3	*11-14/02/08*

The first round of the individual interviews was conducted before the teachers took part in the PD workshop. In addition, the first focus group interview was integrated into the second workshop. The second round of the individual interviews was carried out when I began to provide relevant basic principles of performance-based assessment. At the same time, the participants started to revise the rating criteria. The second focus group interview was integrated into the final workshop. Finally, the last round of the individual interviews was conducted after the final workshop. By the time of the final interviews, the participants had been given the input on fundamental principles of performance-based assessment along with the hands-on experience in

developing rating criteria. For more details of the purposes and processes of each interview, see Section 5.2.2.3.

6.2 Main Study Data Overview

In the following sub-sections, I will describe the overview of the data of each the workshop participants. The data is drawn from three individual interviews and two focus group interviews. The overview of data is divided into four major themes: thinking about assessment, thinking about rating criteria, thinking about PD, and finally reported assessment practice. For the process of the derivation of these themes, see Section 5.3.2.2.

6.2.1 Participant 1: Catbandit

Catbandit is 29 years old and has been teaching at the Department for approximately 6 years. Catbandit teaches linguistic courses for English major students as well as FE courses. He was a coordinator of FE 1 when this study was conducted.

6.2.1.1 Thinking about assessment

In focus group interview 1, Catbandit expressed his view toward the assessment tasks. He thought that they were not authentic, and added that the instructions for the tasks were redundant. In Interview 1, Catbandit pointed out that one of the most important considerations in language assessment was what and how to assess students. He also emphasised that the assessment criteria must directly reflect the syllabus of the course. The weighting of assessment tasks should also correspond to the objectives. In other words, Catbandit was very much concerned with the content validity of the assessment. For the assessment used in the FE courses, Catbandit recognised that there were two types of assessment being used: traditional examination and task/performance-based assessment. He perceived exams as a necessary part of the course assessment because it was the conventional practice of

the Department. In addition, he pointed out that because some aspects could not be assessed by the classroom tasks, those aspects could be assessed by the exam. Catbandit, moreover, reported that he was satisfied with the use of the performance-based assessment in general. However, he was not satisfied with the fact that the assessment did not include the assessment of the process of completing of the task. He commented that:

> *We want students to complete the tasks and we only get the task but we don't get to assess the details ... For example, we teach learning strategies but we don't assess if students can use these strategies. And we don't assess the process of finishing the tasks. It's like we teach them how to complete the tasks, but we don't get to see how well they do it. We only get to see the final finished tasks.*

Catbandit was not satisfied with the content validity of the performance-based assessment used in this particular context because there was no ongoing assessment for the tasks. In consequent, Catbandit questioned the validity of the assessment and whether it truly reflected the ability of students. Because the assessment did not allow him to see the process of completing the tasks, Catbandit questioned if the finished tasks could reflect the ability of students as there were many other factors involved. He reported that:

> *Practically, we can't be sure if students could learn and really do [giving oral presentations] ... And we can't be sure if they use their true ability in completing the tasks ... because there are many factors such as some students read the scripts during the presentation, they didn't write the tasks themselves ... they might have copied them from somewhere else.*

Moreover, Catbandit stressed that there were many assignments to rate, which decreased rating quality. In focus group interview 2, Catbandit supported his view toward the assessment he expressed in the first interview that he did not believe that task-based assessment was suitable for the Department because the majority of

students did not have high proficiency; he believed that this type of assessment did not work with low proficiency students. He reported that most of the students he had taught did not understand the assessment tasks.

In the third interview, Catbandit stated that assessment was very crucial in an educational system, especially for students' learning. Though it was only a part of the system, assessment had been perceived as a synonym of education. Therefore, as a teacher who had to assess students, Catbandit emphasised that teachers had to make sure that assessments were efficient, fair and valid. He said "*I want to take part in making sure that assessment is effective and can truly assess students, and is fair"*. Moreover, in order to improve assessment, Catbandit was aware that teachers needed to have knowledge in assessment. He found it difficult to get involved in improving assessment: *"I haven't learned about the theory of assessment. And I think it is the only reason that I can't do it well enough because I don't know if what I do is right or wrong"*. Furthermore, Catbandit stressed that validity, reliability and fairness were the most important aspects of assessment. He reported that he perceived himself to be a fair rater because he followed the available criteria. He wanted to make sure that the scores are valid and reliable by applying the same standards to every performance. In other words, Catbandit wanted to be intra-rater reliable. He explained that "*I don't want to see students of the same level of ability are awarded different scores. I'm aware that there are many factors affecting the scores I award. So I always go back to the performances I've already rated*". In addition, Catbandit pointed out that rating students' performances could be complicated because there were many factors involved, especially the subjectivity of raters. Nonetheless, he believed that rating criteria could help control these factors or decrease the subjectivity.

6.2.1.2 Thinking about rating criteria

In focus group interview 1, Catbandit pointed out that the descriptors in the criteria were not in the correct domains, and they should be rearranged. When asked about

the rating criteria in Interview 2, Catbandit only commented that prior to the PD he had a little background about language assessment and rating criteria. However, in Interview 3, Catbandit expressed in length his attitude toward the criteria, both the existing and the revised criteria. He made a general comment on rating criteria that they could "*control those factors affecting rating or decrease the subjectivity. They [rating criteria] make it easy for teachers to rate and increase rating consistency, which also lessen our worries*".

When talking about the existing rating criteria, Catbandit pointed out that the activity (when the participants rated the samples of students' performances and shared their opinions about the criteria) allowed him to realise that there were problems with the criteria. He stressed that "*It became clear when we tried to rate the samples of the performances that there were problems with the (existing) criteria*". With the revised criteria, Catbandit still thought that there were some problems which had not been solved because the revision was not completed. He suggested that there was not enough time to pilot and make necessary improvements. Nonetheless, he pointed out that the revised criteria, compared to the existing ones, were clearer and easier for teachers to use.

6.2.1.3 Thinking about professional development programme

In Interview 2, Catbandit emphasised what he liked about the PD was having the opportunity to share ideas and opinions with other participants, especially on the issues relating to the course and its assessment. The only problem he reported was the restricted time, as he himself did not have much time and it was hard to find opportunities when every participant was free. Moreover, Catbandit expressed his positive attitude toward the PD. He stated that the PD had academically provided him with new knowledge and ideas in language assessment, particularly about the rating process and scoring methods. He said:

> *I had no knowledge at all about the existing rating criteria – how they were created and why they were created this way. We just used them because they were there. But from the workshops, I've learned that we should revise the criteria.*

Apart from having direct benefits, Catbandit also reported that the PD was beneficial for the Department as a whole. He stated that the revision of the criteria would help improve the quality of assessment of the course. Furthermore, he hoped that after the criteria had been revised the Department would adopt them into use for the course. Catbandit also went further, and stated that other courses should revise the criteria by following the procedures laid out by the PD. In focus group interview 2, Catbandit stressed that the PD had illustrated the need for changes in every level, from the course materials to assessment. He added that, "*The PD has reflected that we have to create rating criteria which can decrease problems concerning rating. The criteria have to be able to help raters agree with each other as much as possible*". Therefore, he believed that it was important to train teachers regarding assessment as well as materials development.

In the third interview, Catbandit maintained his positive attitude toward the PD. He stressed how much he had learned "*innovative*" ideas and concepts in language assessment, especially on developing rating criteria and the rating process. The PD also gave him chances to put these ideas and concepts into practice. In other words, not only had he learned the theories of how to develop rating criteria, but he also had the opportunity to develop the criteria. In addition, he pointed out the direct benefits of the PD to the Department; that is, the revision and improvement of the assessment criteria of the course. Moreover, Catbandit added that the PD, especially the think-aloud, had increased his awareness of the consistency in rating.

6.2.1.4 Reported assessment practice

In the first interview, Catbandit stated that when he rated students' performances, he followed the rating criteria. However, by the end of the interview, he admitted that because of a heavy workload he did not follow the criteria strictly. He said:

> *Sometimes I can't do it [strictly follow the criteria]. I look at the overall picture and check if it [a student written performance] has included all the required elements... And from experience of many semesters teaching the course, I know the patterns of the performances and can award the scores accordingly.*

He also reported this in focus group interview 1 that when he started working he was very strict with the criteria, but he became less strict when he had more experience. In Interview 2, Catbandit pointed out that despite the rating criteria he had to use his impression when he rated the students' tasks because of his heavy workload. He said that because he had to do the coordinating for the course, as well as teaching, he did not have time to pay attention to all the criteria described in the scales.

In the third interview, similarly to the previous interview, Catbandit stated that previously he did not have *"consistency"* when rating students' performances. He reported that:

> *When I started working as a novice teacher, I strictly followed the criteria. But with more experience, I began to use my experience and impression to rate students' performances. From being strict with the criteria, I became much less strict.*

However, Catbandit emphasised that the PD had changed the way he rated his students' performances. The first reason the PD had an impact on his assessment practice was the use of think-aloud in the research process. Catbandit said that the use of think-aloud helped raise his awareness of the rating criteria while rating the performances in, as well as outside the think-aloud sessions. He stated that though he

only had to rate a few performances for the think-aloud, he had to rate the rest of the performances exactly the same way because he wanted to be "*fair*" to every student. Thus, Catbandit concluded that the use of the think-aloud method made him rate the performances more "*consistently*". He reported that:

> *I could say that I've become aware of rating more consistently. Before I participated in the workshop, there was no consistence when I rated students' performances … After the think-aloud sessions, I had to do the same with other students because it's the same task. I had to use the same approach to make it consistent.*

Catbandit, moreover, defined himself as a "*fair rater*". He clarified that he "*followed the rating criteria*" and he made comparison between the performances he was currently rating with the ones he had already rated to ensure that the same level performances were awarded the same score. In other words, he was concerned with the reliability of his ratings. Furthermore, Catbandit reported that he had to follow the Department's assessment conventions set out by "*previous generation teachers*" because he did not have knowledge in assessment. He also stated that he followed the assessment conventions because they were "*orders*", even though he did not know any rationales behind these conventions.

6.2.1.5 Summary

The central theme derived from Catbandit is the impact of the knowledge he has acquired from the PD on his rating style, attitude towards the assessment and the roles of teachers in assessment. Learning and understanding about performance-based assessment, especially on the rating process, made him become more self-consistent when rating. In addition, Catbandit becomes more critical to the assessment being used. This knowledge in assessment also allows him to be aware of the role teachers can play in regulating the quality of the assessment. The discussion of these themes will be explored in Section 7.1.1.

6.2.2 Participant 2: Papone

Papone is 34 years old. He earned an MA in TEFL. He has been teaching at the Department for approximately 6 years. Papone was a coordinator of the previous fundamental English 1, and has been a coordinator of FE 2 since it was first implemented. He was also one of the material developers of this course.

6.2.2.1 Thinking about assessment

In the first interview, Papone expressed his preference in traditional examinations. He explained that exams could truly assess students' achievement. He said that it was a very "*traditional way*" of thinking but "*practical*". However, Papone was aware of the roles of performance-based assessment, and he added that assessment had to reflect the course's syllabus. When the syllabus included speaking and writing, thus, the assessment had to include these two skills.

With the assessment used in the two FE Courses, Papone expressed his concerns of the lack of a midterm exam and listening test. He said that personally he wanted to have a midterm exam because students could make a decision of dropping or continuing the course when they learned the scores from the exam. In addition, he supported including listening in the assessment. Nevertheless, Papone, as a coordinator of FE 2, stressed that including a midterm exam and listening tests could cause many management problems. He stressed that "*The term management covers many things and influences many decisions*".

About the performance-based assessment used in the courses, Papone was especially concerned with the reliability of raters. He said that:

> *The number of sections that teachers teach in each semester affects how they rate the performances. So if they teach many sections, they have to rate many performances. For example, some teachers might rate 30 performances consecutively. I mean some teachers might be able to handle it but some teachers aren't aware that they can't. So for these teachers, after having rated 20 papers, the scores they award might become unreliable. But it doesn't apply to every teacher especially expert teachers. It's really different from one teacher to another.*

Though he was aware of this problem, Papone pointed out that the use of performance-based assessment in the course was a good practice and it was the trend of language assessment. However, he recognised that it increased workload for teachers in rating students' performances.

During Interview 2, Papone pointed out that he became aware of some problems of the assessment, and added that these problems were caused by carelessness of the material writers (including himself) during the development process. He clarified that there were some aspects which were not assessed but should otherwise be assessed, and vice versa, and there were some aspects which were not assessed well enough. Papone also said that this problem was reflected in the rating criteria.

In the third interview Papone maintained that he and other material writers overlooked assessment, especially the criteria and scales, when developing the course materials. They focused more on designing the syllabus and the tasks. After having participated in the PD, Papone realised that:

> *We overlooked some aspects at that time. Now I wonder why we don't assess these aspects. For example, we teach and review the use of 'which' and 'where' but they don't appear in the criteria. And we aren't serious with 'nice layout' but it appears in the criteria.*

In other words, he becomes aware of the mistakes in the assessment he created.

6.2.2.2 Thinking about rating criteria

In focus group interview 1, Papone, who developed this rating scale, thought that the rating criteria were generally acceptable. He also added that a holistic scale, which was the method employed by the scale under investigation, was suitable for experienced teachers. In the first interview, Papone pointed out that within the same course, different tasks used different scoring methods which may cause confusion among teachers. However, he stated that the criteria were "*rules*" that teachers must follow. Papone added that the criteria were set up by the material developers and they expected teachers to follow them. Papone justified that he did not propose this because he was the coordinator and one of the material writers. He stressed that when he taught other courses he followed the criteria very strictly: "*When I teach other courses, like FE 3 and FE 4 of which I'm not a coordinator, I follow all the criteria and guidelines. Some courses require teachers to do very tedious arithmetic and I follow them*". Moreover, Papone was aware that training could help improve the quality of rating, but he was not sure if it could eliminate the problems because "*eventually we can't check if teachers follow the criteria (after training).*"

In Interview 2, Papone became aware of problems with the criteria and scales. One of the problems was that the criteria did not include necessary aspects. This problem was caused by the carelessness of the material developers. He stated that:

> *In the rating criteria, we might not have included some aspects, which we teach in class and should be assessed, in the criteria. Some aspects in the criteria don't assess students well enough. And we have included some aspects in the criteria which shouldn't be there at all.*

In the second focus group interview, Papone added that the problems with the criteria were caused by the fact that the criteria were holistic. In the third interview, Papone pointed out that the development process of the assessment (including the criteria and

scales) of the courses did not include interactions or dialogue among the material writers because of time pressure. The course materials and assessment were developed by a group of teachers who individually worked on their own chapters. They were supervised by the chief advisor who oversaw four courses which were being developed at the same time. After each teacher finished his or her own chapter, the materials were distributed to senior teachers for comments and feedback. Papone added that there was no formal meeting among these teachers and the materials writers. He stressed that there was not any dialogue, such as in the PD workshop. Furthermore, Papone stated that they did not have this kind of meeting because there was a time pressure among the materials writers.

With the revised criteria and scale, Papone thought they were easy to use compared to the existing ones. Nevertheless, he said that it did not mean that they were without flaws. The most obvious advantage of the revised analytic scale, was the reliability of the score, Papone added. He said that the revised scale included clear descriptors of each level of performances compared to the existing scales, thus the scores derived from the revised criteria should be more reliable. However, Papone pointed out that when he first saw the scale (complied by the researcher) he was shocked because of its look. His first impression was that the scale was very detailed compared to the existing scales. Nevertheless, Papone stressed that when he tried to use the scale, he did not have any problem with it, and he thought it was "*ok*". He added that the scale should be piloted and revised if it was going to be used in the course. Papone also reported that when he rated students' performances in other courses, Papone began to feel that the scales were not clear. He stated that "*Sometimes I feel ... perhaps I might be thinking about the PD workshop and feeling that the tasks and the criteria of those courses don't have any depth or something like that*".

6.2.2.3 Thinking about professional development programme

In Interview 2, Papone said that he participated in the PD because he wanted to learn about language assessment. He said he took some courses in language testing and assessment about 10 years ago. Thus, he realised that participating in the PD would be a brushing-up activity for him in this respect. He was also aware that he would learn new concepts. Papone pointed out that he would also be able to get involved in practical works, as most of what he learned in his previous courses were very much theories. After having participated in the PD for a few sessions, Papone stated he had gained new experiences, especially using think-aloud as a research method and rating process. He stated that "*I've learned new ideas about assessment particularly in rating process, especially developing rating criteria and scales*".

Furthermore, Papone stressed that the PD had created "*new perspectives*" for him. He said that in the past he (and other material writers) assumed that "*This is the way everything should be like, for example, the criteria and scales, but the PD has shown me the alternatives*". In addition, Papone recognised one major difference between his previous experience in developing assessment and the experience in the PD. He reported that when he developed the assessment for FE 2, he and the team did not study if the assessment was relevant to the course objectives or course syllabus, whereas he did in the PD. Therefore, he strongly believed that the PD had helped him analyse the relevancy of the assessment and the course. With the revision of the criteria, so far, Papone thought that the criteria had become more relevant to the objectives of the courses. Nonetheless, Papone did not think that the PD had made a drastic change to the criteria. He believed that the participants in the PD had re-organised the criteria to make them easier for teachers to follow. He stressed that this was to make the criteria less complicated and the scores to be more "*valid*". Apart from new perspectives, Papone emphasised that he had learned about assessment from the PD, particularly on criteria and scale development process, and rating

process. The process of this learning included the input from the researcher, as well as sharing ideas with the participants while trying to revise the criteria.

In the second focus group interview, Papone stressed that the PD had changed his view toward the criteria. He stated that "*Some aspects, in the criteria, I didn't notice or I didn't care that they were problems. They weren't in my head before. But when we, in the workshops, investigated the criteria, I realised that some criteria shouldn't be there. The workshops have changed my perspectives.*" In Interview 3, Papone maintained similar views toward the PD. He stressed that the PD had broaden his perspectives and raised his awareness of the present problems of the assessment. In addition, Papone described that he had learned new ideas as well as about the participants' opinions and ideas through discussions. He hoped that he would be able to implement what he had learned from the PD in the future. Furthermore, Papone reported that he decided to participate in the FE 4 criteria revision project led by the FE courses advisor.

6.2.2.4 Reported assessment practice

Papone did not report his assessment practice in the interviews. He maintained in all the interviews that he followed the criteria very strictly.

6.2.2.5 Summary

The PD has given Papone "*new perspectives*" in assessment which allows him to critique the past, present and future of his assessment practice. The knowledge and experiences he has acquired made him aware that the process of which the criteria were developed was one of the causes of problems of the rating criteria in the Department. In addition, Papone is planning to apply what he has acquired in the near future. The discussion of these themes will be explored in Section 7.1.2.

6.2.3 Participant 3: Tanya

Tanya is 32 years old. She is the least experienced teacher in the group. She has taught for approximately 5 months. She has an MA in TEFL, and also in Applied Linguistics. Tanya started her teaching career as a part-time instructor and later was granted a full-time position, just before the second individual interview.

6.2.3.1 Thinking about assessment

In the first interview, Tanya expressed that her ideal language assessment would be an on-going assessment (or formative assessment) of which it was not under the exam condition but being done in the classroom. She said she wanted to be able to see students' development:

> *I feel that when we assess students, we don't have to use test like a final exam. I want to assess from the development of the students, from the activities they do or their homework... I prefer this kind of assessment. And I don't want students to feel that they're under the exam condition. Assessment has to be an on-going process. It's the responsibility of teachers to assess students during the term.*

However, she accepted that in reality (in this context), classroom size made it impractical. In order to do this kind of assessment, Tanya believed that teachers had to devote a great deal of time for each student and it was a lot of work. She stressed that an examination was more practical for 6,000 students because it was easy to mark. She also believed that exam was a necessary component of assessment because it could actually assess students' achievement of what they had learned, especially grammar and vocabulary. She pointed out that the exam could assess student's competence and understanding of vocabulary better than performance-based assessment could:

> *Sometimes, students know the meaning of the words but they can't use it in the context. When they have to fill the words in the blanks (cloze test), they just can't do it... But they can when they have to use the words to give a presentation ... It implies that they don't really understand the meaning of those words*.

Nevertheless, Tanya viewed a performance-based assessment as the assessment of performance, whereas an exam as the assessment of competence. She recognised that performance was not the direct reflection of competence. She said that some students might be better at competence whereas some students at performance. Therefore, Tanya believed that there should be both a "*standardised*" test and performance-based assessment. In addition, marking exams was easier for teachers. With performance-based assessment, she said that it should include various tasks. For speaking assessment, she preferred the assessment in which students had to speak spontaneously without notes, for example, an interview by the teacher. For writing assessment, Tanya added that the assessment should consist of different kinds of prompts.

With the assessment used in the FE courses, Tanya thought it was generally acceptable, but she felt that the weighting of the final exam was too high and the tasks for performance-based assessment were not authentic. However, she expressed that the performance-based assessment was similar to her ideal assessment. She could learn about students' ability from their task's performances and be able to give them support when needed. She stated that, at least, she could give support to those students who had problems with the tasks they performed during the semester. It would be too late to find out after the final exam, she added. Though this increased the amount of work for teachers, Tanya thought it was the responsibility of teachers. She strongly agreed with the on-going way of assessing students. Tanya concluded that:

> *I love seeing students' level of ability from the beginning. So I know if they have any problems from the tasks they submit... I know that there will be more work for teachers. I feel that I totally agree that assessment has to be an on-going process. I like it. It's teachers' responsibility.*

In Interview2, Tanya expressed her view toward the weighting of the assessment; particularly that she did not agree the final exam should receive the highest weight. She said that because the course was a task-based course, thus, the tasks should weigh highest. In addition, she said that performance, or language, should be the main focus of the assessment, not other aspects of the task such as presentation skills. In the assessment tasks for the FE courses, the majority of the tasks were done in groups. Tanya pointed out that there should be more individual tasks. Concerning group work, Tanya said that she wanted to mix high and low proficiency students because they could help each other, but she was aware that sometimes high proficiency students would do all the work. Furthermore, Tanya stressed that the objectives and the criteria for the assessment of the tasks must be clearer. She also suggested that students should keep all the tasks and produce a profile, or portfolio, of the performances. Finally, Tanya, similarly to the previous interview, maintained that she would like the assessment to be an on-going process because "*I can see students' development. I don't like the fact that students have to take exams. In the exams, students' roles are very passive as they are aware that they are under the exam condition. I really don't like it*".

In the third interview, Tanya described that because the course employs a task-based syllabus, the assessment criteria should focus on language more than other components. She said "*I don't know if this domain should weigh less than other domains. I think the domain 'language patterns used' should focus on accuracy... It's a task-based course. I don't know if we can make this component (content/task fulfilment) weight less*". Tanya also expanded her views toward the assessment of

other courses. She expressed that she was "*very worried*" with the rating. She thought there were problems with the assessment, especially on rating process. She said:

> *I have to admit that there are gaps (problems) with rating. When I rate the assessment tasks, I feel that we don't have standards. I feel that rating depends on how we feel at a time of rating. Personally, I try to rate all the papers the same way but the problems are the temperament and fatigue. Sometimes I get very tired after rating many papers*.

In other words, Tanya began to recognise problems relating to the rating process of other courses, especially on the reliability of raters.

6.2.3.2 Thinking about rating criteria

In the first focus group interview, Tanya, as a new teacher, pointed out that she needed a rating scale with clear descriptors. She suggested that the exiting criteria were not clear enough as she needed a scale comprised of a checklist in which she could tick off the required aspects. Tanya added that she could not rate using her impression because she was an inexperienced teacher. In Interview 1, Tanya emphasised that she had problems with following the criteria, though she did not state that the problem was with the criteria. She said that "*Sometimes I can't make a decision where the performance is on the scale*". In Interview 2, however, Tanya pointed out that before she participated in the PD she had thought that the criteria and scales were appropriate and good enough for the course. She reported that:

> *Because the criteria and scales were prepared by the Department, I thought they had been carefully designed. They were appropriate for the tasks. But when I actually use them, I've found problems. First I thought it was my problem that I didn't understand the criteria or that I couldn't make decisions. But I've just found out that it wasn't the case.*

She stated that it was from participating in the PD that she realised it was the criteria that caused the problems. Tanya described that the think-aloud activity, and

discussing with other participants, raised this awareness. She said "*In fact, there are many weaknesses which need to be improved. It's like thinking it aloud with other participants and I've found that my thinking is similar to others'. And in reality, it's the problem of the criteria which has consequences*". The problems Tanya discovered included unclear and overlapping descriptors. She reported that "*In the scale there is a criterion which should be separated into two different criteria... I can't judge them as one criterion*". The example she cited was the criterion for rating the delivery of a presentation:

> *For example, I feel that a student gives a presentation very naturally though he sometimes looks at the script. But the rating criteria state that he could only get half of the full score. And pronunciation is included in this criterion. In some cases, students have good pronunciation but they often look at the scripts. So what should I do?*

In Interview 3, Tanya maintained that she had problems with the holistic rating scale (she did not use the term holistic scale in the first two interviews). She pointed out the weakness of holistic scales used in the course that the descriptors were too broad and without clear directions. She stated that:

> *A holistic scale to me is like a blank page... I'm not good at giving holistic scores. I think in some criteria, the descriptors are very broad. There are different problems with different students. There isn't any explanation or direction in the scales.*

Tanya also added that holistic scales allowed awarding efforts, and she reported that she sometimes awarded extra marks for students.

In contrast, Tanya had different views toward the revised scales. She said that the revised scales had clearer directions with more detail, which would help rating. She thought that the revised analytic scales had more control over teachers' rating which, in consequence, would make scores more valid. Tanya stated that:

> *The scales are more detailed than and not as broad as the existing ones – which help with rating. Though there might still be some problems, of course the criteria don't cover everything yet, at least they help with my decision making when I do the rating. I prefer specific descriptors. They control how I award scores.*

Nevertheless, Tanya recognised that the next step was to prepare teachers to use the scales. She stressed that "*We must train raters – train them how to make decisions based on the criteria*".

6.2.3.3 Thinking about professional development programme

In Interview 2, Tanya expressed that she participated in the PD because she wanted to continue learning, that she only recently graduated and it was the first year in her teaching career. She said "*I wanted to take part in anything that would help develop myself … to learn about different aspects of teaching. I want to acquire new knowledge to develop myself*". Because she was a new teacher, Tanya said that she had some questions and problems, which she was not sure if it was just her or other teachers did as well. This PD, therefore, was the opportunity for her to discover that other teachers also had the same problems. Tanya added that discovering problems led to improvement. She pointed out that without the discussions with the participants, she would not have realised that it was not her that was the problem, but the criteria and scales.

Furthermore, Tanya expressed her views from what she had learned about assessment from the PD. She stated that she had learned about the rating process from sharing opinions and experiences in rating with other participants. Tanya also added that she had learned about research. She stated that before the PD she had thought that research in language testing and assessment was all about scores and quantitative research. She had learned from the PD that research in this area could be a qualitative research with fewer participants. Moreover, Tanya emphasised that she particularly

liked the think-aloud technique. In interview 2, Tanya reinforced her previous views toward the PD and the rating criteria. She recognized that before the PD, she could not follow the scales which were holistic. She said that because she did not have experience the criteria, it did not make sense to her. Therefore, she had to manage and create her own scales. However, she pointed out that after the PD ratings became clearer to her. She stated that "*The workshops have made me more confident when I make judgements about students' performances. I feel that I do that based on principles*". Tanya also stressed that as a new teacher she did not feel confident in other debates. She pointed out that she felt she had developed in terms of her thinking about assessment from taking part in the PD.

In the third interview, Tanya stressed that the criteria and scales developed in the PD derived from the problems in the context:

> *What we have created originated from the problems. We weren't led by any propaganda or anything. We had the problems and we talked about it. We didn't set up what it had to be. But we worked based on the principles. To solve the problems, we had to base on testing principles. Eventually, what we've achieved resulted from our discussions.*

Tanya was aware that the Department provided supports for teachers' development. However, she commented that the Department should provide more in-service training for teachers. She stressed that "*If it's possible, there should be meetings to brainstorm ideas to revise the criteria of other courses and for other tasks of FE 2*".

Tanya pointed out that the PD gave her opportunities to scrutinise her rating style and other participants', especially from the discussions on think-aloud protocols. She stated that the think-aloud made her "*aware of my own rating style*". In addition, she said that she was particularly keen on the activity in the PD when the participants rated students' performances and shared their opinions on the criteria. Tanya liked this activity because she had the opportunity to share her rating style as well as learn

other teachers' rating styles. Finally, Tanya reported that the PD had made her become more confident in making judgements and rating. She said "*The PD workshop has made me become much more confident, especially when I award scores. I dare to, for example, this performance shouldn't get this mark. I might have become more severe*". Tanya pointed out that she had higher expectations from students. She explained that before the PD, she considered students' background as one of the criteria. For instance, she would have lower expectations from students majoring in Science than languages. However, after having participated in the PD, Tanya reported that she did not consider the background of the students but focused on the performances. Tanya added that she felt she had more "*consistence within myself*" when she rated students' performances.

6.2.3.4 Reported assessment practice

In Interview 1, Tanya reported that she was not comfortable with the existing scales, so she created her own version of the scale on a spreadsheet (MSExcel). In the spreadsheet, she listed all the criteria to create a check-list to use when rating. She stressed that "*Some students might have some aspects but don't have others. I can't judge a performance from an overall point of view. I just can't*". When she awarded a score, she would rank the performances before assigning a score. She reported that "*I look at the best and worst performances first. Then I look at that performance and decide what level that performance fits in*". Another problem Tanya found when rating was that the performances were not good enough. To solve the problem, Tanya returned those performances with feedback and asked students to revise them. She added that she would award the revised performances not the ones with problems.

In the second interview, Tanya maintained that she could not use holistic scales. She emphasised that "*The existing scales, I just can't award scores using holistic scales*". Similarly, in Interview 3, as described previously, Tanya pointed out that a holistic scoring method allowed awarding efforts. Tanya awarded extra points

for efforts when she rated the performances using holistic scales. She said that "*Some students have very low proficiency. I feel pity for them. So when I rate their performance, I gave them extra points for their efforts, which isn't in the criteria*".

In order to achieve consistency in her own rating, Tanya reported that after having rated many performances, she would go back to the ones she had already rated to remind herself how she rated them. She thought that it was time consuming, and it was difficult to keep consistency. She finally added that when there were clear and standard guidelines in rating, she would strictly follow them.

6.2.3.5 Summary

The PD has provided Tanya with opportunities to deconstruct and understand her rating style. When she becomes aware of her rating style, she establishes her way of rating her students' performances. This process also leads Tanya to become confident and self-consistent in her rating. The discussion of these themes will be explored in Section 7.1.3.

6.2.4 Participant 4: Wanwisa

Wanwisa is 34 years old. She has been teaching in the Department since 1996. She has an MA in Applied Linguistics. Wanwisa was one of the material developers of FE 4 (English for Humanities and Social Sciences) as well as the coordinator for this course.

6.2.4.1 Thinking about assessment

In the first interview, Wanwisa expressed that in a skill-based course, she preferred the assessment which used "*bands*" and an assessment method which consisted of the components of "*performance*" and "*competence*". She reported that previous courses were competence-based in which assessment and teaching focused too much on grammar and reading. She believed that they did not allow students to show their performance. The results were that many students dropped off and failed the courses.

In contrast to these courses, Wanwisa pointed out that though the present FE courses were lacking grammar teaching, they consisted of performance-based assessment, which encouraged students to take risks in using the language. She also stressed that these courses gave the students opportunities to demonstrate their levels of ability which, in consequence, promoted a positive attitude toward learning English. Wanwisa believed that performance (as opposed to competence) was what students needed in real life situations. Furthermore, Wanwisa believed that teachers were also motivated because with performance-based assessment, they engaged in the act of teaching more. She stated that:

> *I like these courses. In the previous meeting, many teachers agreed that though these courses are weak in terms of grammar teaching ... but what we've seen was students don't hate English courses the way they used to. They have better attitude toward English. They have the courage to speak, take risks and make mistakes ... I prefer these courses to the previous ones.*

However, Wanwisa was aware that there were problems with the ratings of the courses. She pointed out that there were factors affecting how teachers awarded the scores. She stated that "*Because with this type of assessment, there is no right or wrong answer, we might not be consistent. And having to rate many performances is another factor*". In addition, she reported that having more than 30 students in one class was another factor affecting rating: "*Especially teachers who teach 3 sections, I have to admit that they aren't consistent*".

In Interview 2, Wanwisa expressed her awareness of the significant role assessment played in education. She described that in many courses, regardless of teaching approach, it ended up focused on grades. She felt that a great deal of attention was paid to grades; therefore, assessment had to be appropriate with what students learned. Furthermore, Wanwisa pointed out that assessment, apart from being suitable for students, had to be accessible for teachers. She emphasised that

teachers had to be able to manage assessment with the class size of more than 30 students. In other words, assessment had to be "*teacher friendly*". In addition, teachers should be provided with a rating process of which they felt "*secure*". Wanwisa said that "*There should be ways to make teachers feel certain with clear conscious when rating. And teachers must feel secure. They don't have to be worried that someone will question their ratings*".

Furthermore, Wanwisa reported that from having scrutinised the FE courses in the PD, she began to think about the assessment of the English major courses. She felt that in English major courses, assessment should cover other aspects rather than language alone. She said that:

> *With the English major courses, we have to think about quality. In order to compete in the competitive labour market, only language isn't enough. Our students have to have quality. So we have to include other aspects in the assessment.*

For the FE courses she expressed that it was very difficult to assess the quality aspects because these courses required students to produce certain quantity components of language rather than the quality.

In the past, Wanwisa believed that assessment in the Department was fixed and could not be changed. Therefore, she thought "*Why pay any attention because everything was fixed*". She added that "*Nobody stood up and said anything about assessment, Foundation courses or English major courses*". However, Wanwisa had recently noticed that there could be changes. She stated that "*When you (the researcher) started talking about assessment in the Department, some senior teachers agreed with you. So I think we can change the criteria for other courses in the future*".

In Interview 3, Wanwisa expressed her concerns over performance-based assessment. Her main concern was on the reliability which she believed to affect students' grades. She said that:

> *I'm worried with the FE courses because there might be discrepancy between sections. Though there are criteria, we don't know how each individual teacher were apply them. And there are weaknesses and problems with the criteria. I think these factors do affect students' grades, especially these FE courses.*

Nevertheless, Wanwisa was not worried when she had to develop items for examinations. She said that she followed the conventions: "*There are fixed rules. I only follow them so I'm not at all worried. There are specific conventions*". However, she was not satisfied with the exams used in the FE courses. She thought they comprised of confusing items. Overall, she said the exams did not meet "*standards*" and was not sure if the exams were developed on any testing principles. Wanwisa stated that "*Exams, we don't have any standards. I don't know if they've been developed based on any theories. It seems like they've been done based on the coordinators to fit the required marks, 56 points for FE 2*".

Wanwisa also thought that the exams were "*unfair*", especially the answer keys. She reported that the answer keys were too restricted to what was taught in class. When students produced the answers taught in class, they were not credited. In addition, she was not satisfied with the proportion of item types and the weighting system being employed in the exams. Wanwisa stated that:

> *The answer keys are too restricted to what we teach in class. But sometimes students' answers are acceptable for communicative purposes. In the answer keys, we can't accept these answers because we don't teach them this way… And what criteria do we use in writing the exams and the weighting? For example, in a reading comprehension part, a true/false item weights only half a point. Who set up that vocabulary items should get one point and context clue half a point?… I feel that writing exams based on teachers' needs isn't fair for students.*

Moreover, Wanwisa believed that how the Electronic-Self Access Language Learning (E-SALL) operated was not fair. She reported that students were not given the criteria for the assessment of the E-SALL. She said that "*Last year, we told students that they would get the full mark if they had 700 correct items. But this year we don't tell them how many because we were worried that students would cheat... It isn't fair*". In addition, Wanwisa did not agree with the use of E-SALL as part of the assessment because it was too demanding as students had to do the exercises online, and get 800 correct items in order to get a full mark. Also students had to do three quizzes in class, of which items were taken from the E-SALL exercise. Wanwisa felt that it was too demanding. She stressed that "*We assigned all of these rules without considering students. We don't feel sympathy for the students at all*".

6.2.4.2 Thinking about rating criteria

In focus group interview 1, Wanwisa expressed that her attitude was indifferent, compared to other teachers. This was because she had taught the course many times; it had become more like convention. She also pointed out, from her experience, that following the criteria and using her impression resulted in similar scores. Thus, she thought that impression marking was much easier than following the criteria. In Interview 1, Wanwisa expressed that for her, the criteria were only "*guidelines*". She also explained that the existing criteria allowed differences in rating from teacher to teacher because "*There aren't any clear descriptors which discriminate students'*

levels". She reported that if she followed the criteria, performances at level 3 could be fitted in the level 4 on the scales. However, she did not believe that this inconsistency would affect students' grades. She stated that "*Though the rating isn't 100% consistent, the differences aren't that great. It doesn't make a C+ student get an A or an A student get a B, for example*". She added that teachers shouldn't let the criteria completely dominate their ratings. Teachers must use their own judgements when rating.

In the second interview, Wanwisa maintained her view that the criteria were not appropriate and needed revision. She described that:

> *Some criteria aren't appropriate because they are too detailed. They have been used for a while but the Department hasn't analysed the questions the teachers have raised. They should study why some teachers don't like the criteria or their weaknesses and strengths. There are pros and cons but they haven't been discussed. It seems like the criteria have been set up to stay.*

Furthermore, Wanwisa reported that she began to recognise the weaknesses of the criteria of other courses. She pointed out that she had seen the criteria for a writing course for English major students and felt that there were problems, as the criteria and scales were too detailed. Wanwisa also reported that she had told the coordinator of that specific course to talk to the researcher for ideas to improve the criteria and scales for the course.

Moreover, similarly to the previous interview, Wanwisa pointed out that the rating criteria were merely guidelines for teachers. When rating, "*We need to use our own judgement along with the criteria. For example, in the scales, the criteria describe the requirements of the task, when rating we need to consider these requirements together with our own judgement"*.

In Interview 3, Wanwisa pointed out that the existing criteria focused too much on the quantity, not the quality of the language. She expressed that the criteria

should include aspects of quality in the descriptors. Apart from the quality of the language, the scales themselves had to meet a certain set of standards because she believed that the quality of the criteria affected students' grades. Wanwisa stated that "*There are problems with the criteria and they affect students' grades*".

In terms of different types of scoring methods, though Wanwisa stated that she personally preferred a holistic scoring method to an analytic one, she pointed out that analytic scales helped teachers become more consistent, but, she did not think that either of them was easier to use than the other. Nevertheless, she pointed out that for analytic scales, there should not be too many criteria. She said that "*With the revised (analytic) scale, I wouldn't agree if there would be more criteria… We have to limit to 4 or 5 criteria. I don't agree with 6 or 7*". Wanwisa referred to a writing course of which its scales comprised of detailed rating scales, and expressed that "*I'm so glad that I don't teach this course anymore. The rating criteria for this course are so tedious. I don't agree with them at all*".

With the revised criteria from the PD, Wanwisa stressed that in order to implement them, it was very crucial that other teachers also understood them. She said that teachers needed to be well informed about the revised criteria in detail. She pointed out that they might not understand the revised criteria the same way the participants did. She said "*At least we need to explain to teachers that the criteria have been well studied and developed with appropriate principles. If they understand this, I think they will follow the criteria when rating*".

6.2.4.3 Thinking about professional development programme

In the second interview, Wanwisa reported that the Department did not provide any in-service training similar to the PD. She recognised the significance of the dialogue among teachers in the PD. She stated that "*Our Department doesn't often have this kind of professional development. It's like each teacher does what they are good at. There isn't any interaction or exchange of knowledge among them. There is no*

collaboration among teachers". Wanwisa also pointed out that it was important to include participants from different backgrounds in the PD, especially teachers with more experience. In the present study, she recognised that Songsri was a resourceful person in the team. She stated that:

> *It's good that we have Songsri in the team because we can have opinions or views from teacher from another generation. Songsri has experiences in developing courses… It's good that we have one experienced teacher. If we only had junior teachers, I think it wouldn't have been this good.*

Furthermore, in her opinion, a PD should be informal: "*Usually I don't like assessment. But I've found the PD fun and beneficial... I like that it isn't too academic... I don't like when it's too formal and academic. I like personal and affective approach*".

Wanwisa described that she had learned from the PD how ratings should be done. In addition, she had begun applying this new learning to other courses. She said that "*In other courses that I'm teaching, I've begun to mark the homework by creating my own scales*". Wanwisa also believed that in the future, for new courses, she could apply the experience gained from the PD to the assessment of these courses. She added that the researcher should start similar projects with other courses. She said "*I feel that assessment of some courses aren't appropriate … I'd like you (the researcher) to approach other courses (to do similar projects), especially English major courses*". Wanwisa believed that the assessment of these courses could be improved, and there were possibilities that the Department could do the improvement, because some senior teachers had agreed with the researcher about revising the criteria and scales of the FE courses. For the present project, she hoped that the revised criteria and scales would be put into use for the course. She stated that "*What we're trying to accomplish is the revised criteria. I hope that they'll be*

implemented. And I hope that the revised criteria will be suitable for our teaching and they'll be accepted".

Wanwisa also reported that, in the past, she did not voice her opinions about assessment because she felt that she did not have the authority. However, after having participated in the PD, she began to express her opinions. She stated that:

> *In the past, I didn't talk and speak about assessment because I felt that I didn't have the authority. But at least, you're (the researcher) doing a PhD in assessment and have experiences or doing research and have read about what others in the field have done – you now have the authority. And I'm telling you these (her ideas and opinions) because what you've said is similar to my ideas.*

In focus group interview 2, Wanwisa argued that the problems in ratings were caused by the lack of training for teachers – not because of the type of scales used. She stated that "*If we don't have training, be it analytic or holistic scales, there will always be problems*". She also stressed that the PD had helped her realise the strengths and weaknesses of the assessment tasks which could lead to improvement. She stated that the PD had helped her realise the significant roles of the criteria. She admitted that before the PD, she relied on her experiences when rating, not the given criteria. She explained that she did not follow the criteria because sometimes the criteria did not make sense to her.

In Interview 3, Wanwisa reported that from the PD, she had learned about the rating process and became aware of the weaknesses of the course's assessment. She said that:

> *I've learned that what we need to include in the assessment and the acceptable assessment practices. And I've become aware of the weaknesses of the assessment we're using. You know, I didn't think about these before. They came up from the workshops.*

Wanwisa pointed out that she would like other teachers to participate in these kinds of activities because she wanted junior teachers to have the opportunity to learn. She emphasised that these activities should be similar to the ones in the PD.

Furthermore, Wanwisa pointed out that, from sharing and brainstorming ideas with participants, who were from different backgrounds, she gained new perspectives about assessment and became aware of the problems. She declared that these problems had been brought to light by the PD. She stated that:

> *Before the workshops, we might be aware of these problems but they were not clear. Or we might be aware of them but we just ignored them… But now we're aware that these problems will affect students and their learning. And we know that changes are possible. So I feel that there must be changes.*

Though Wanwisa admitted that she preferred a holistic scoring method, she agreed that the revised analytic scale in the PD should be implemented. She pointed out that because there were weaknesses with the existing scales, the revised analytic scale should be best in this situation. Moreover, she expressed that the revised criteria should be put into use because she was part of the revision team. She said "*Because I got involved in revising the criteria from the beginning, I'm part of the team, so I think, as one of the participants, we should continue (implement the scales). And other teachers should participate too*".

6.2.4.4 Reported assessment practice

In the first interview, as stated previously, for Wanwisa the criteria were only her "*guidelines*". She reported that when she rated students' performances she used her "*conscience*" and "*experience*" as her rating framework. Moreover, Wanwisa considered "*creativity*" as one of the criteria, though it was not stated in the scales. She said that "*I also consider creativity. For example, a student who produces a correct task exactly as the model might get the same score as a student who makes some mistakes but his task is creative and different from the model*".

In Interview 2, Wanwisa reported her experience rating a writing course for English major students. She was strict with the criteria but was later informed by the coordinator that she did not have to be so strict. She reported that:

> *It was the first time I taught that writing course. The criteria were very detailed. I followed them. The criteria were very tedious ... But when I told the coordinator that I had problems following them because they were too detailed, the coordinator told me not to be too serious with them. The criteria were only guidelines ... But I followed them anyway ... However, if I ever to teach this course again, I won't follow the criteria. I won't waste my time. I'll assign a score holistically.*

In the third interview, Wanwisa maintained that when she rated students' performances in the FE courses, she considered "*creativity*" as one of the criteria. She did not pay much attention to "*grammar*". In addition, when students submitted the tasks late, she did not deduct any marks, as described in the criteria. She stated that:

> *I gave students opportunities to present their ideas and I allowed grammatical mistakes as long as they use the language. So I don't mind if they make mistakes. I focus on creativity – I mean something new and interesting... And I'm not a severe rater. For example, when they submit their tasks late, I don't deduct any marks because I haven't rated them yet. I don't see any point of deducting one point as other teachers do.*

Wanwisa also added that she viewed students' language from a "*holistic point of view*". She said that "*I tend to view how students use the language in a bigger picture rather than a discrete point of view … For instance, for a writing task, if I think it's ok overall, communicative wise, I wouldn't look into detail if there are any run-ons or fragments*". However, Wanwisa reported that when she taught English major courses she would do it differently. She explained that because English major students had more language input and more exposure to the language, she would have higher expectations in grammatical accuracy from these students. Furthermore, when

employing the rating scales, which consisted of separate criteria for "*language*" and "*content*", Wanwisa reported that she could not distinguish between the two. She usually ended up awarding the same score for both domains.

Wanwisa also reported that the persons in charge of the FE courses (i.e. the advisors and coordinators) had an important role in influencing how assessment would be conducted in the Department. She stated that:

> *Concerning assessment in the Department, it's like whoever is the coordinator or the advisor, we have to follow their orders though some of them are impractical or unreasonable. We have to because they have authority... I don't deny that I follow what the advisor has told me as well.*

She added that experience was also a crucial factor affecting her interest in assessment. In the past, she did not pay attention to assessment much. However, having taught in the Department for more than 10 years, she began to question how assessment had been employed.

6.2.4.5 Summary

The data reveals the gradual changes of Wanwisa's attitudes toward the rating criteria, the roles of teachers in performance-based assessment, and her behaviour in rating. Prior to participating in the PD, Wanwisa thinks that teachers should rely mainly on their judgements and use criteria only as guidelines when rating students' performances. However, after having participated in the PD, she gradually realised that teachers have to follow the criteria to make rating more consistent. Moreover, Wanwisa has been trying to follow the criteria as well. The discussion of these themes will be explored in Section 7.1.4.

6.2.5 Participant 5: Songsri

Songsri is 49 years old. She is the most experienced teacher among the participants. She received MA in Applied Linguistics (English for Science and Technology) in

1993, the same year she started teaching at the Department. Songsri was one of FE 4 (English for Science and Technology) material developers. She also had considerable experience in test development and writing.

6.2.5.1 Thinking about assessment

In the first interview, Songsri expressed her view toward language assessment in general stating that various methods should be used in assessing students' language ability. However, she did not like traditional testing, especially a cloze test because it was too difficult. She stressed that examinations should not trick students and be too difficult, as they could discourage learning. Assessment should motivate and encourage students' learning. Songsri believed that explicit and clear objectives and criteria were one of the most important qualities of assessment, in addition to the validity.

For the assessment of the FE 2, Songsri thought that they were generally acceptable. However, she did not agree with the delivery of the final examination. She thought that the exam was too difficult. Songsri pointed out that the reading passages, and the item types, were too hard for the students. She believed that the format of the exam was too mechanical. She said:

> *The item writers relied on the item writing conventions which caused the exam to be too difficult. This is what I strongly disagree with … The item writers forgot that students might only be able to handle discrete items only. But it seems that we're proud of writing long discourse items with, for example, complex reading passages. Some vocabulary items are tested where students have to complete a well-written passage with given words. They have to do two things, reading comprehension for the whole passage and vocabulary. Many came up with zero!*

Songsri also emphasised that though she did not agree with the exam, she could not do anything about it. She stated "*The exam is where I don't have any authority. There's no way that I can have power over it*". She reported that she had a conflict

with the exam writing committee. She was one of the committee and responsible for one part of a reading section. One of the committee members, who was the "*expert in testing*", did not agree with a passage she chose for her part and the items she wrote, because they were too easy. Nevertheless, as Songsri reported, they were used in the exam.

Furthermore, Songsri stated that the tasks for the assessment were not in the correct order of difficulty, and the criteria as well as the objectives for the tasks were not clear. Songsri believed that clear criteria could motivate students' learning because the tasks with clear criteria, for students, were achievable. She also added that teachers played very important roles in the assessment. Therefore, teachers had to be flexible and adequately understand the assessment. Songsri said that "*Teachers don't have to follow every step in the Teacher's Guide ... Teachers have to be able to adapt the lessons ... And teachers have to know the right moments to emphasis about assessment*". She emphasised that in a classroom context she had the power over assessment, whereas she did not have any power in the exam situation.

In the second interview, Songsri maintained that the assessment tasks were not in the correct order of difficulty and they were not authentic. In addition, she believed that assessment should have appropriate criteria, based on local context and students, and consist of clear and explicit objectives. She stated that:

> *We have to think about our students' levels. We can't expect 'native' or 'near native' level or use bands developed for native speakers. So we have to set up the criteria with our students, who are non-native speakers, at the centre ... and from our context... And we must not use us, the teachers, as a standard, but instead base it on one first year students' ability.*

Songsri, in addition, disagreed with failing students because she believed that it was discouraging. She believed that failing students did not benefit anyone. She said that ELT at a university level should encourage students not scare them.

In Interview 3, similarly to the previous interview, Songsri stressed that assessment had to be context specific, by being based on students' point of views and needs. Songsri maintained that the examination was too difficult. She stated that the exam writing convention was unfriendly to students, and the contexts of the reading passages used were too alienated. Thus, Songsri concluded that the exam was discouraging, unfair and not valid:

> *This morning I looked at Student's Workbook to check if we include 'possessive form' in the lessons because it is tested in the exam. No, we don't, but we test it in the exam. I question if this is fair for students. The exam team didn't check if we test what we teach... God – it's too hard to earn. It's too difficult. The exam isn't friendly for students. It isn't fair.*

Similarly to the previous interview, Songsri described that though she wanted to make changes to the exam, she could not do anything, because she did not have any authority to do so. She stated:

> *In the committee, sometimes I just can't argue with them because the majority of them are senior teachers who have authority. If they decide to do it this way, we can't do anything. So I think this (the PD) is the way to educate them. Well, I wouldn't call it educating but sharing. It's the easiest way, I think.*

Nevertheless, Songsri stressed that in her own classes and the course she coordinated, she felt that she had authority and could arrange the assessment the way she believed they should be.

Furthermore, Songsri pointed out that the assessment did not have clear and explicit objectives, which she believed led to validity questions. She also emphasised that a lot of problems in assessment were caused by the fact that teachers did not have sufficient knowledge in assessment. Finally, Songsri added that she believed that ELT

in Thailand had not been successful, because, we did not teach what students wanted to learn and the tests were too difficult. She also stated that:

> *We aren't aware of students' context; their needs and wants … Do we teach what is relevant to their needs? Do we only teach what they want? … Sometimes we test them on what we haven't taught… and the tests are so difficult that no one can do it, only the teachers can, of course. Yes, because teachers are clever so they choose difficult items. It isn't fair.*

6.2.5.2 Thinking about rating criteria

In focus group interview 1, Songsri argued that she did not agree with how the descriptors were grouped in the existing criteria. She pointed out that some do not belong together, for example, the descriptor 'speak fluently – not reading the script all the time' should not be put under the domain 'Language' but 'Presentation skills' which was not one of the criteria. She added that the descriptors were grouped this way for the convenience of the developers. In the first interview, Songsri emphasised that the existing criteria were not specific enough. For instance, one of the criteria required students to use 'emotive adjectives', but the descriptor did not clearly state how many emotive adjectives students needed to use in order to achieve the mark for each level. She argued that the criteria must specify the number of emotive adjectives for each level of the performances. Songsri believed that students could get practice using the adjectives this way.

In Interview 2, Songsri added that one of the problems with criteria was how different teachers applied them. She said that when rating students' performances, some teachers were too severe because they wanted to comply with the "*standards*". These standards, she believed, were imposed on them by other teachers, especially those who teach English major courses. She said:

> *Teachers are scared of not having standards which have been set up by some teachers here. And, these standards have been forced on students. I don't know if we're trying to protect our own necks or promote learning. Many teachers have become too scared of these standards.*

In the third interview, Songsri maintained that counting was one of the best options, though participants in the PD did not agree (as it was discussed in the PD whether the criteria should specify the number of 'emotive adjectives' for each level in that criterion). She said that including counting in the criteria would encourage students, because students would feel that the task was achievable. She stated that:

> *I still argue that when we teach lexical items, for example, 10 words, we need to know how many words students can use to be called 'excellent'... There must be this kind of criterion because students have different ability. So those with higher ability can perform better.*

Nevertheless, she reported that with the revised criteria and scale, she felt "*more comfortable*" because the criteria consisted of clearer descriptors and the scale was based on "*assessment principles*".

6.2.5.3 Thinking about professional development programme

In the second interview, Songsri, based on her psychology background, viewed the PD workshop as a behaviour modification. She recognised that in the workshop, "*The participants share what they are and absorb from others. And finally we achieve something in the middle*". Songsri did not think that she had learned anything new from the PD. She said the PD had confirmed about assessment theories she already knew and learned; she stated "*I can't say that I've learned anything new. I wouldn't say something innovative*".

However, Songsri recognized that the PD was a learning opportunity for the participants, including her. She said that in the PD, the participants could learn about Western and up-to-date concepts in assessment, as well as research methodology

from the researcher. Songsri pointed out that this PD was not only for the benefit of the research per se, but it would benefit the Department as a whole. In addition, the participants, who were the "*key persons*" in the Department, would be able to put into practice what they had learned in the future, for instance, in writing new courses. Songsri later clarified that these key persons were teachers who could influence changes, for example, the coordinators. Furthermore, she recognised that because the participants were from different backgrounds, and not all of them had similar educational backgrounds, the PD was the opportunity for them to learn from each other. It was, especially, the opportunity for junior teachers to gain experience. She stresses that "*It's already been good. It's (the PD) for the junior staff to learn*."

In this learning process, Songsri added that she became more aware about criteria and scale development process. Nevertheless, Songsri thought that the PD could be improved by including a couple of senior teachers. She also stressed that the PD should be reported to the Department and expanded to other courses. In the second focus group interview, Songsri pointed out that the PD had reinforced her views, that it was important to include teachers who were responsible for material development in a PD because they had to be aware that materials and assessment had to go together. She explained that one of the main reasons the participants took part in this PD was to help the researcher; thus, she believed that to make the materials writers aware of assessment issues, the policy makers had to introduce a policy which aimed at mandating material writers to take part in a PD in assessment.

In Interview 3, Songsri reinforced what she said in Interview 2, that the PD raised the awareness in assessment among the participants. She recognised that "*Everyone became aware of the important elements of assessment*" and "*We've arrived at the consensus point together*". Moreover, in the process of revising the criteria and scales, the participants learned many aspects of assessment, especially assessment criteria. As Songsri had realised the importance of the PD, she suggested

that the PD should be expanded to other teachers and courses to improve assessment. She also recognised the researcher as "*the person who has had a lot of input*", and therefore was in the position to carry out the PD in assessment for other teachers in the Department. She expressed her desire that the researcher should make other teachers aware of how they should conduct assessment. The researcher should also set up assessment practice guidelines and report them to the Department.

6.2.5.4 Reported assessment practice

In the first interview, Songsri thought that the criteria for the assessment were not clear. For example, they should specify the exact number of emotive adjectives students were required to use to fulfil the task. She, thus, made changes to the criteria by telling the students the number of adjectives they had to use in the task. She reported that "*Because we have set up the objective that students have to use emotive adjectives in this task ... so when I teach I emphasise it. I tell my students that they have to use at least 3 emotive adjectives in this paragraph*".

In Interview 3, Songsri reported her past assessment practices and compared them with her present practices. In the past, Songsri was very strict with grammar when rating students' performances. She said "*I used the bible written by native speakers and applied it with Thai students. For example, if they didn't have the correct word order when forming a question, I gave them zero*". However, in the present time, she reported that "*I tend not to be very strict with grammatical rules. If they don't have the correct word order when forming a question but the sentence makes sense, I only deduct a few points depending on the communicative quality*". Songsri explained that this change was due to having witnessed the failure of English language education in Thailand.

6.2.5.5 Summary

It seems that Songsri believes that her knowledge and skills in assessment are already aligned with the core principles advocated in the PD workshop. Therefore, she does not find it necessary to make any improvement. Likewise she maintains that the PD has not affected her thinking and her practices. The discussion of Songsri's resistance to change will be further explored in Section 7.2.3.

6.3 Confirmatory Study: Findings from the Follow-up Study

The follow-up study was carried out between June and July 2007, three months after the main study. The aim of the follow-up study was to further investigate the impact of the PD on the teachers who participated in. In the phase, the first interview (Interview 1) was carried out before the participants applied the revised rating criteria when rating students' performances, and the second interview (Interview 2) after they had rated the performances. In this section, I explain the research design and the overview of the data.

6.3.1 Research design

The follow-up study was conducted as a confirmatory study. The below sub-sections include the purposes of the study, its research questions, and its data collection process.

6.3.1.1 Purposes of the study

In order to confirm the findings from the main study, in the follow-up study, the objectives were to:

1 Study whether the teachers, who participated in the PD, have further changed

2 Further investigate the effects of the PD on the Department assessment practices

3 Further explore the assessment policies in the Department by focusing on how they have affected the teachers in the focus group and those with responsibility in assessment

6.3.1.2 Research questions

The following research questions were explored in the phase of the study:

1 Have the teachers who participated in the programme changed? If yes, in what way? If no, why not?

2 In what ways were other teachers affected by the PD programme?

3 How has the assessment culture in the Department affected both groups of teachers?

6.3.1.3 Data collection process

After the main study, I left the research site for three months, which was a summer holiday when the majority of teachers were working on revising or developing materials. When I returned to the Department, I found out that the Department had decided to revise the criteria for the FE courses. At the time that I arrived, they had already revised the criteria for FE 1 based on the criteria developed for FE 2 in the PD. However, because it was the beginning of the semester, the criteria were not used yet. Therefore, I planned my interviews into 2 rounds: before the participants used the criteria to rate their students' performances in order to examine the teachers' attitudes towards the criteria, and, after they had used the criteria in order to investigate whether their attitudes have changed. (For guiding questions, see Appendix K.)

Interview 1

The focus of the interviews was to obtain insight into assessment beliefs, knowledge, and attitudes of the teachers who participated in the PD. The interviews focused on the teachers' views of the assessment tasks, assessment process, assessment products, and the new rating criteria that the FE 1 committee had recently developed to be

implemented that semester. The main purpose of this round of interviews was to compare its findings with the interviews on the same topics in the main study, and to investigate if there are any changes in the participating teachers. Moreover, the questions also covered their views toward the PD and how they thought the PD had affected them.

Interview 2

The aim of this round of interviews was to elicit the participants' views toward the rating criteria after they had used them. The participants were also asked to compare the criteria with the ones they developed in the PD. Moreover, the interview questions covered how they rated the written tasks using the new criteria. The data from the interviews were used to be compared with the data from the main study, to explore changes in the participants.

6.3.1.4 Participant profiles

The participants in the follow-up study are the same as in the main study. For their profile, see Section 5.2.2.4.

6.3.2 Data overview

Following the main study, in presenting the data from the follow-up study I firstly present the overview of the interviews. The data includes four themes: thinking about assessment, thinking about rating criteria and thinking about the PD (though they are not separated into different sections).

6.3.2.1 Participant 1: Catbandit

In terms of his views towards the assessment, in the first interview, Catbandit did not think that task-based assessment is appropriate for this context because students did not understand or appreciate this approach. He stated that "*Students don't understand the tasks. They don't understand the process of a task-based course… Some students*

don't pay attention to the process… Sometimes, students just copied the tasks from somewhere else." Thus, Catbandit thought that the Department should use different methods by combining different methods.

Concerning rating criteria, in Interview 1, Catbandit questioned the changes of the scale from 4 to 5 levels and the examination from 56% to 50%. He pointed out that the changes were not based on any empirical study or principles, but came from the agreement of teachers in the meetings. He also stressed the difficulty in including level 5 in the scale, as he was part of the revision team. Nonetheless, he thought that the revised criteria should work well because the descriptors were clearer. He believed that the revised criteria could increase the rater reliability. After the revised criteria were implemented and he had used them in rating students' performances, Catbandit reported in Interview 2 that though he had found that the criteria were easy to follow, there were performances that did not fit in any descriptor in the criteria. In addition, he thought that the majority of teachers still did not fully understand how to rate the criteria using the new rating criteria. He stated that "*The teachers who strictly followed the criteria wouldn't have many problems. But teachers who used their impressions would. For example, they might feel that the scores derived from the criteria are higher or lower than their impression. These teachers would have and cause problems*".

As far as the PD is concerned, in Interview 1, Catbandit pointed out that the PD was very beneficial because it provided him with knowledge which he had actually put into practice. He stated that he applied what he had learned about criteria development in the criteria revision for FE 1. Catbandit also reinforced in the second interview that he tried to follow the steps in revising the criteria he had learned from the PD when he helped revise the criteria.

In conclusion, although he thinks that the revised criteria are easy to follow because of the clear descriptors, Catbandit believes there are some teachers who do

not strictly follow the criteria and still use their overall impressions when rating students' performances. Furthermore, Catbandit criticises that the decisions on the changes to the assessment were based purely on teachers' intuitions. He argues that these decisions should be based on assessment principles or empirical studies.

6.3.2.2 Participant 2: Papone

Concerning assessment, in Interview 1, Papone expressed his views that he did not agree that the Department should change the assessment of the courses based on what teachers agreed upon in the meetings. He believed that changes should be derived from empirical studies or principles. He stated that "*The changes shouldn't just happen like that (from the meetings). We must start with questions of why we need to change. We have to study the problems of what we have. It's very important*". He also pointed out that the changes the Department had adopted did not solve the problems from the core. He suggested that "*We have to start from analysing the course objectives to the assessment*". Papone, moreover, stressed that he was not sure if the changes would create positive or negative impact.

As far as the revised rating criteria are concerned, in the first interview, Papone, as part of the criteria revision committee, reported that he found it was very difficult in creating and using level 5 for the criteria. He also pointed out, from taking part in the revision, that the existing criteria were not appropriately developed. He stated that he and the team did not take the objectives of the tasks and courses into consideration. Thus, there were many problems. After having used the revised criteria in the course, Papone reported in Interview 2 that he was satisfied with the criteria for the written task, though there were some problems with the criteria for oral presentations.

In terms of the PD, in Interview 1, Papone stressed that the revised criteria (from the revision project) followed the criteria developed in the PD. He also reported that he used the procedures in revising the criteria as in the PD. Thus, he concluded

that the PD had raised his awareness in assessment, especially the rating process and the process in developing rating criteria for performance-based assessment. He said "*It (the PD) is very beneficial. I've learned also about assessment. I've paid more attention in assessment – something that I previously overlooked… I hope that other teachers would become more aware of the significance of assessment*". He also added that it was the PD and the researcher that triggered the changes in assessment in the Department. He pointed out that "*It's you who created these changes… There must be someone to endorse the changes, directly or indirectly*".

To summarise, Papone does not agree that the Department decided to revise the rating criteria because some changes to the assessment were based purely on teachers' opinions in the meetings. He argues that these decisions should be based on assessment principles or empirical studies. In addition, Papone has been following the steps in developing rating criteria he acquired from the PD when he revised other sets of rating criteria for the course.

6.3.2.3 Participant 3: Tanya

Concerning her views towards the revised rating criteria, in Interview 1, Tanya reported that she believed that the revision of the rating criteria would help increase the reliability because she believed that the existing criteria were not clear. She reported that the new revised criteria would make assessment fairer. She said "*The criteria act like guidelines telling teachers to rate with a certain set of standards. But teachers have to make their own decision in some cases*". Nevertheless, in Interview 2, after having used the criteria in rating students' performances, Tanya reported that there were some problems with the criteria. She said that the descriptors did not cover all possible performances. Furthermore, she pointed out that some teachers had told her that they had deducted marks when they felt that the scores were too high. Tanya admitted that she had a similar view. She also felt that the scores derived from using the revised criteria, which were analytic scale, could be higher than the scores derived

from the previous holistic rating scale. Tanya pondered that "*I don't know when we follow the criteria [which employ analytic rating method], do we have the rights to deduct the scores according to our feelings… Maybe, I think we need to revise the criteria*".

In terms of the PD, in Interview 1, Tanya, as a member of the revision staff, reported that she used the revised criteria from the PD as a model in the revision project. Tanya also pointed out that the PD also helped her to understand more about rating criteria. In the meetings during the revision, she was able to explain to other revision staff some principles. In Interview 2, Tanya realised the significance of rating a sufficient number of the samples of students' performances in a rater training. She pointed out that the recent rater training did not present different levels of performances; thus, in the actual ratings, she had problems when the performances were not represented in the training, and they were not covered by the descriptors.

In summary, the knowledge and experiences Tanya gained from the PD have helped her understand about the rating process. She has applied what she has learned in the revision of the rating criteria for FE 1. Furthermore, this enables her to become aware of the shortcomings of the revised criteria and the rater training.

6.3.2.4 Participant 4: Wanwisa

In terms of her views towards assessment, in the first interview, Wanwisa stressed that in a task-based syllabus, when the emphasis was on performances, the assessment must not focus on examinations. However, the exam accounted for 50% (56% before the revision) of the assessment, which Wanwisa considered too much. She stated:

> *Tasks are what we focus on in the courses. They are continuous process… and the outcome of the process (i.e. the finished task performances) is what we expect students to perform. This is what we expect, so it should have the highest weight in the assessment.*

She also added that, because of this high weight of the exam, the final (letter) grades did not tell how well students performed on the tasks. Thus, she pointed out that she agreed with adding band 5 to the criteria because she believed that students deserved higher weight for the performance tasks. She said that because students had put a lot of work to the tasks, the tasks should have higher weight.

Furthermore, with the changes of the assessment in the Department, Wanwisa pointed out that she was aware that changes would cause confusion among teachers, but she stressed that they should accept the changes because they aimed for improvements. However, she pointed out that, at that moment, the Department should not make any further changes, but focus on how to improve the current situation. The problems with the current assessment were students cheating and the inappropriate format and items in the exam. Wanwisa reported that the exam focused too much on vocabulary and some items were too tricky and difficult.

Concerning rating criteria, in Interview 1, Wanwisa emphasised that regardless of teaching methods, whether task-based or grammar-based, the criteria must be appropriate. In Interview 2, after having used the revised criteria in rating students' performances, Wanwisa pointed out that she did not agree with the revised criteria and that the criterion 'content' be weighted more than other criteria. She also stressed that the criteria 'content' and 'language' should be under the same criteria. Furthermore, Wanwisa added that, though there were new rating criteria, she thinks that many teachers still don't follow them but still heavily relied on their own impressions.

As far as the PD is concerned, in the first interview, Wanwisa describes that the PD was a good thing because it gave her opportunities to share, discuss, and talk about the problems of assessment, which created new perspectives for her, especially in ratings. She adds that from sharing ideas with participants in the PD workshop, she became aware of what to take into consideration when doing her own ratings.

Furthermore, Wanwisa points out that the PD led to the improvement of the course's assessment at large.

To conclude, Wanwisa thinks that there have been enough changes to the assessment, and, the focus should be now on improving what is in place. In addition, Wanwisa does not think it is valid that the revised criteria focus too much on the content. She also thinks that many teachers still do not strictly follow the criteria.

6.3.2.5 Participant 5: Songsri

In Interview 1, Songsri expressed that the change of the weight for the task, which she considered as minor change, from 4 to 5, was a positive change because students deserved higher score for the performance tasks. However, major changes were not possible because of the policy of the Department. Teachers who are in power could not implement changes because they did not have the necessary knowledge to create changes. Thus, Songsri believed that the Department needed collaboration among teachers to make changes. She emphasised that politics played very important part in the success of creating changes. She stated:

> *Sometimes politics is very important. If we aren't part of the community, for example exam committee, we don't have any rights or power to argue with them. They tend to only respect the opinion from the committee. So I think we must change this practice… It's like who is in power has the rights and authority to justify his or her decisions.*

For the assessment used in the courses, Songsri thought that they were acceptable, but the problem was the course materials. Nevertheless, she reported that some aspects of the assessment needed improvement. For instance, the assessment did not include listening skills and the tasks were not authentic for Thai students. For her, assessment must be fair and relate directly to the objectives of the course.

Concerning rating criteria, in the first interview, Songsri expressed her reaction toward the change of the criteria by adding band 5 and that it would not have

much impact on how teachers did the rating. She pointed out that it was the teachers who played the most important role in rating. Teachers consciously compared the performances of students between the two levels. In Interview 2, after having used the criteria, Songsri described that the revised criteria were clear and easy to use, which helped teachers in making judgement on students' performances, though she reported that she did not agree with one criterion. Songsri also pointed out that by nature teachers would follow the criteria but each had a different style.

In Interview 1, Songsri described the PD as the opportunity to share with the participants her frustration concerning the courses and their assessment. She also added that it provided the chance for participants to make improvement to the assessments. She pointed out that, in the PD, the researcher had the role of providing the knowledge in assessment. From the discussions, the participants became aware of problems in assessment, especially its unfair aspects and the problem of content validity. She stated that:

> *You helped us by giving the knowledge that we didn't know. We, as practitioners, realised that the assessment didn't work. We punished the students… And we were not fair with them… We then became aware that we must test what we teach.*

In Interview 2, after having used the criteria, Songsri reported that she did not have any difficulty in following the criteria because she had already been 'sensitised' with the criteria in the PD. She reported that she was satisfied with these new criteria because they followed the criteria developed in the PD, and she felt good being part of constructing the new criteria.

In summary, Songsri maintains that the PD does not have any impact on her. She claims that the PD is for other participants to learn, as she did not learn many new ideas from it. Furthermore, she reinforces that the PD is beneficial to the participants and the Department as a whole.

6.3.3 Summary

In this section, I have presented the overview of the data from the follow-up study. The data confirms that the PD has had positive impact on the participants, except one participant, Songsri, who remains unchanged. The data indicates that the participants have applied the knowledge and experiences they have acquired from the PD workshop in their assessment practices as well as their other responsibilities in assessment. Further discussion on the follow-up study and the main study on the impact of the PD on the participants' rating styles and attitude toward the assessment will be presented in the following chapter.

6.4 Conclusion

In this chapter I have presented the data overview from the main study and the follow-up study. Table 6.3 below summarises the categories and codes derived from the analysis of the data. The data, from the categories and codes, reveals that most of the participants have been affected by their participation in the PD workshop. The impact of the PD on each participant will be presented in Chapter 7 Section 7.1.

Table 6.3 Summary of categories and codes

Catbandit		
Interview 1	*Interview 2*	*Interview 3*

Assessment
• Course
  o Exam
  o Management
  o Objectives
  o Task performance-based
• General
  o Course objectives
  o Validity
• Workload: Rating: Decrease quality
Practice
• Criteria
  o Follow
  o Not follow
	Assessment
	• Rating: Based on criteria
	Criteria
	• Lack understanding
	PD
	• Benefits
	  o Academic purposes
	  o For many people
	  o Learning
	• Expectations
	  o Other courses
	  o Real applications
	• Impacts
	  o Criteria
	  o Learning
	  o Revise the criteria
	• Strengths: Discussions
	• Weaknesses: Time constraint
	Practice
	• Depend on criteria
	• Past
	  o Didn't question
	  o Followed
	  o Lack of knowledge
		Assessment
		• Factors affecting rating
		• Fairness
		  o Comparing performances
		  o Following the criteria
		• In education
		  o Crucial
		  o Overarching
		• Literacy
		  o Necessary
		  o To improve assessment
		• Role of teachers
		  o Make it efficient
		  o Make it fair
		  o Make it valid
		• Validity
		  o Criteria
		  o Scores
		Criteria
		• Increase reliability
		• Original: Found problems
		• Revised
		  o Clearer & Easier to use
		  o Not completed
		  o Some problems

PD

- Benefits
  - Experience
  - Improvement
  - Innovation
  - Learning
  - Revise the criteria
- Impacts
  - Awareness in rating
  - Consistency
  - No impact
  - Think aloud
- Weaknesses: Time constraint

Practice

- Comparing students
- Consistency
- Criteria
  - Follow
  - New criteria
- Detailed
- Follow conventions
  - No expertise
  - No rationale
  - Orders
- Holistic & Analytic
- Inconsistency
  - Novice vs. Experienced
  - Think aloud

Papone		
Interview 1	*Interview 2*	*Interview 3*

Assessment
• Course
  o Management issues
  o Midterm
  o More workload
  o Performance-based
Criteria
• Differences
• Rules: Ts must follow
• Training could help
• Practice
• Criteria
  o Strictly follow
  o Study before rate
	Assessment
	• Weaknesses
	  o Carelessness
	  o Missed some important aspects
	Authority
	• Include advisors: Increase reliability of criteria
	Criteria
	• Original: Don't cover necessary aspects
	PD
	• Benefits
	  o Academic purposes
	  o Aware of weaknesses
	  o Experience
	  o Learn new knowledge
	  o Refresh past knowledge
	• Impacts
	  o Aware of the weakness of course assessment
	  o New perspectives
	  o To improve the criteria
	• Strengths: Discussion
		Assessment
		• Course assessment: Weakness
		Criteria
		• Original
		  o Development process
		  o Irrelevant aspects included
		• Other courses
		  o Not clear
		• Revised
		  o Easier to use
		  o First impression
		  o Increase reliability
		  o Should be piloted
		PD
		• Benefits
		  o Broaden perspectives
		  o Future application
		  o Learning others' perspective
		• Impacts
		  o Aware of weaknesses
		  o Future application
		  o New perspectives
		  o New project
		  o Question other courses' criteria
		Practice
		• Awarding efforts
		• Follow criteria strictly

- Weaknesses
  - More participants
  - Senior participants

Tanya

Interview 1	*Interview 2*	*Interview 3*
Assessment • Course o Exam o Task performance based • General o Combination o Communicative test o Ideal: Assessment for Learning o Reality: Exam o Performance o Traditional exam **Criteria** • Expand from the original • Her way makes it less subjective • Problem with levels **Practice** • Ask Ss to revise • Can't give overall scores • Expand from the original • Follow instructions • Rank performances • Use spread sheets	**Assessment** • Course o E-SALL: Lessen the weight o Dislike exam o On-going: Ss' development o Task-based **PD** • Benefits o Learning o New teacher o Problem solving o Share ideas • Expectations o Opportunities to share ideas o Revise the criteria • Impacts o Become more careful o Criteria o Think aloud o Want to improve	**Assessment** • Other courses o Criteria too board o Fatigue o Unsure • Rating o Worried • Task-based o Should focus on language **Criteria** • Need training for Ts • Not happy with the scores o Scores from holistic vs. analytic • Original (Holistic) o Allow awarding efforts o Had many questions o Just followed them o No clear directions o Ss are different o Too board • Revised (Analytic) o Different weight for domains o More controlled o Scores are more valid

	Practice • Holistic scoring: Unable to follow	o Clear directions o Help rating o Might be some problems left o More detail • Supports o Lack for new teachers o To provide trainings **PD** • Benefits o Discover problems o Lead to improvement o Think aloud • Department: Provide in-service • Expectations o Expand PD to other tasks and other courses o Revise the criteria • Impacts o Confident o Higher expectations o Intra-rater reliability • Strengths o Do rating with others o From the problems o Originality **Practice** • Experience o Higher expectations from Ss o More standard

		• Follow instructions
		• Rating
		  o Asked colleagues
		  o Awarding efforts
		  o Comparing
		  o Slow
		  o Try to keep standards
		• Ss' learning
		  o Providing feedback
		  o Special attention to Ss' problems

Wanwisa		
Interview 1	*Interview 2*	*Interview 3*

Assessment
• Course
  o Criteria
  o Factors affecting rating
  o Performance-based
  o Rating
  o Previous courses: Competence-base
• General
  o Prefer band descriptors
  o Skill based
Practice: Rating
• Comparing performances
• Criteria are only guideline
• From experience
• Rewarding creativity
• Setting benchmark
• Use her conscience
	Assessment
	• Course
	  o Couldn't be changed
	  o Hard to assess quality
	  o Nobody asked any question
	• General
	  o Must be appropriate with Ss' learning
	  o Rating
	  o Should be T friendly
	  o Should focus on others than language
	  o Ts should feel secure
	  o Very important in education
		Assessment
		• E-SALL: Not fair
		• Exam
		  o Confusing items
		  o No exam specs
		  o Unfair
		• Exam writing
		  o Fixed rules
		  o Not worried
		  o Follow the conventions
		• Expectations: groups different expectations
		• Performance assessment
		  o Criteria: Quality
		  o Reliability
		Authority: Follow advisor
		• Impractical
		• Unfair

Authority

- Started to listen to the researcher: Possible changes

Criteria

- Other courses
  - Need revision
  - Not appropriate

Department

- Lack of PD activities

PD

- Benefits
  - Learning
  - Professional development
  - Sharing ideas
- Expectations
  - Expand to other courses
  - Implement the criteria
  - Revise the criteria
- Impacts
  - Aware of weaknesses of the criteria
  - Changes of the course's assessment
  - Look at other courses
  - Question criteria of other courses
  - Voice her opinions in assessment
- Unreasonable

Criteria

- Original: Too much on quantity
- Quality: Some aspects aren't appropriate
- Revised
  - Clearer
  - Easy to use
- Scoring methods
  - Analytic VS Holistic
- Too detailed: Disagree
- Well explained: Ts need to be well explained

PD

- Learning: Assessment: Rating process
- New perspectives
- Share ideas
- Expand to all FE courses
  - Co' need to participate
  - Participants from different BG
- To gain experience
- Impacts
  - Awareness
  - Move to analytic
  - Try to be stricter with criteria

Practice

- Follow authority
  - Might be unreasonable & impractical
- Questioning
  - Didn't question

- Strengths
  - Beneficial and fun
  - Informal
  - Participants from different BGs

Practice

- Didn't pay like assessment
- Didn't voice opinions
- Experience in a writing course: Followed the criteria very strictly
- Will use impression
  - Experience
- Rating
  - Allow grammatical mistakes
  - Can't separate language and content
  - Criteria as guidelines
  - Focus on creativity
  - Focus on grammar when teach English major
  - No deduction for late submission
  - Try to be stricter with criteria
  - View performance from holistic point of view

Songsri

Interview 1

Assessment

- Course
  - Exam
  - Generally OK
  - Motivation
  - Not clear
  - Roles of teachers
  - Tasks: Problem: Level of difficulty
- General
  - Exam: discouraging
  - Motivation: Clear criteria
  - Objectives: Most important
  - Validity

Interview 2

Assessment

- Failing Ss: Discouraging
- Her course
  - Appropriate criteria
  - Based on local context
  - Based on students
  - Clear & explicit objectives
  - Well connected
- Tasks
  - Level of difficulty
  - Not authentic
- Washback

Interview 3

Assessment

- Context specific
  - Learners' point of view
  - Students' levels
- Course
  - Exam
  - Objectives
  - Tasks
  - Too high expectations
- Objective
- Ts' lack of knowledge
- Validity

Authority

- In her own class
  - Criteria
  - Feedback
- Powerless: Testing experts: Exam writing
- To be listened to

Criteria

- Not specific enough

Practice

- Criteria
  - Adds her own criteria
  - Make them explicit
- Exam: Use familiar reading topics
- Feedback

Authority

- PD: To include senior Ts
- Standard setting: Imposes on Ts' rating

Criteria

- For accountability
- Too high expectations

PD

- Benefits
  - Behaviour modification
  - Confirming what already knew
  - Improvement to the courses
  - Learning
  - Sharing ideas
- Expectations: Report to the Department
- Strengths
  - From different BGs
  - Involved key persons
  - Learning opportunity
  - Productive & constructive
- Weaknesses
  - Not innovative
  - To include senior Ts

Authority

- Powerful
  - Committee
  - Own course
- Powerless: Committee with senior Ts

Criteria

- Easy to use
- Feel more comfortable
- To specify numbers

PD

- Benefits
- Participants
- Expectations
- Expand to other Ts
- Improvement of other courses
- Participants agreed on the revisions
- Report to the Dept
- Strengths
- Weaknesses

Practice

- Communicative
- Counting
  - Different from PD
  - For Ss' benefits
- Exam writing
  - Achievement
  - Only few hard items

From the analysis of the data, as presented above, three main issues pertaining to teacher change emerged: change in behaviours, attitudes and knowledge. These three issues will be discussed in detail in Chapter 7 Section 7.2. Table 6.4 below illustrates these issues. In behavioural change, the data indicates that the majority of the participants have changed their rating styles. After having participated in the PD, the participants have tried to increase the reliability of their ratings by following the rating criteria. That is, they have become more intra-rater reliable or self-consistent in ratings. Furthermore, it appears that these participants have also changed their attitudes toward assessment. First of all, they have become more critical toward the rating criteria implemented in the Department as they have become aware of the weaknesses and problems of these criteria. In addition, the participants have become aware of the roles teachers play in assessment and the impact of teachers' rating behaviours on students. The data points out that the participants have realised their roles as assessors of their students' performances and assessment developers. They have also become aware of the impact of these roles on students. Consequently, this awareness has affected their rating styles. Thirdly, some participants have become critical toward changes in assessment that took place during the course of the present study.

The final issue that emerged from the data is change in knowledge. Grounded in the data, the participants have learned from the PD about performance-based assessment, particularly in rating process. The data indicates that the participants have acquired knowledge of components of rating process, especially, about rating criteria and raters. The data also reveals that this newly acquired knowledge in assessment has affected their attitudes and behaviours in assessment. In other words, knowledge, attitude and behaviour are interrelated.

Table 6.4: Change in behaviours, attitudes and knowledge

Change in behaviours	
Rating styles: following rating criteria to increase rater reliability	
• sometimes followed the criteria but sometimes did not	Catbandit's Interview 1

• followed the criteria
  o to increase the reliability
  o to increase validity
	Catbandit's Interview 3

• used the criteria as guidelines
• relied on
  o experience
  o conscience
• teachers should rely on their judgements when rate students' performances
	Wanwisa's Interview 2

• tried to follow the criteria more strictly
  o aware that rating affected students
  o to increase reliability
	Wanwisa's Interview 3

• could not follow the criteria • created her own scales	Tanya's Interview 1
• aware of problems (criteria) • adopted analytic scoring methods	Tanya's Interview 2

• experiences in the PD
  o discussions
  o think-aloud
  o ratings
• established her rating style
  o confident
  o self-consistent
	Tanya's Interview 3

Change in attitudes	
Attitudes toward rating criteria: being aware of problems	

• though there were some inconsistency
  o criteria are rules
  o teachers must follow the criteria strictly
	Papone's Interview 1

• there are some weaknesses • caused during development process	Papone's Interview 2
• weaknesses and problems with the criteria of the course and other courses • need revision to increase reliability	Papone's Interview 3
• cannot follow the criteria • new teacher with limited experience	Tanya's Interview 1

• Holistic scale:
  o not clear
  o problems
  o need improvement
• Analytic scale
  o clear
  o helps rating process
  o scores more valid
	Tanya's Interview 3

• assessment has to be valid and reliable	Catbandit's Interivew 1
• questioned the quality of the criteria • did not have sufficient knowledge	Catbandit's Interivew 2

• assessment (including criteria) have to be fair, valid and reliable	Catbandit's Interivew 3
Attitudes toward teachers and assessment: realising roles of teachers in assessment	
• teachers' subjectivity did not affect students' grade	Wanwisa's Interview 1
• teachers had to rely on their own judgements when rating students' performances • teachers should use criteria were guidelines	Wanwisa's Interview 2
• teachers' inconsistency affected students' learning and grades • teachers had to follow the criteria	Wanwisa's Interview 3
• roles as assessor o follow the criteria • roles as assessment developer o make certain the quality of the assessment tasks ▪ reliability ▪ validity ▪ fairness	Catbandit's Interview 3
Attitudes toward change in assessment: being critical to the changes	
• assessment could not be changed • no one ever questioned assessment	Wanwisa's Interview 2
• senior teachers started to be aware of assessment (problems) • revision projects initiated	Wanwisa's Interview 3
• enough changes • should focus on improving current situation	Wanwisa's Interivew (follow-up)
• change should base on empirical studies	Catbandit's Interview (follow-up)
• change should begin with problems and solutions	Papone's Interview (follow-up)
Change in knowledge	
Knowledge of performance-based assessment: acquiring knowledge about rating criteria and raters	
• lack of knowledge	Catbandit's Interview 1
• learning o rating process o rating criteria • not sufficient	Catbandit's Interview 2
• learning and experiences o aware of his rating style o follow the criteria	Catbandit's Interview 3
• new teacher • limited knowledge and experience	Tanya's Interview 1
• learning and experiences o problems with the criteria o improve the criteria o aware of her own rating style	Tanya's Interview 3
• learning o new perspectives o rating process	Papone's Interview 2
• learning o revise criteria for other tasks and other courses o future applications	Papone's Interview 3

7 Investigating Teacher Change: Discussion

This chapter is the discussion of the data presented in the previous chapter. The first part of the chapter traces the changes of four teachers, whose changes were a result from participating in the PD workshop. In the second part, I examine the impact of the PD workshop on these participants. In addition, I offer a possible explanation of the resistance to change of one participant. The final part covers the assessment practices of the English Department. As described in Section 5.3.2.2, that the analysis of the data is guided by Grounded Theory, the discussions in this chapter were drawn from the common themes that emerged from the interview data with the aid of the computer software NVivo. (For the details of the process of how I coded the data, and how I derived the codes used for the interpretations and discussions of the data; Section 5.3.2.2.) In the first part of this chapter, I briefly demonstrate how the themes discussed for each participant were derived by using the maps of the codes created by NVivo. In demonstrating this process, I use an example from the participant Catbandit to show the process in deriving at a conclusion of his change in rating style (discussed in Section 7.1.1.1 below).

Following Corbin and Strauss' (2008) steps of integrating categories, from a rigorous analysis of Catbandit's raw data, as well as the codes and categories derived from the data, I chose 'change in rating style' as one of the 'central' or 'core' categories or themes, with the aid of the NVivo maps. This core theme was drawn from his attitudes towards assessment, rating criteria, and the PD workshop, as well as his reported practice in assessment. Figures 7.1, 7.2 and 7.3 below are samples of the NVivo maps (from three interviews) to illustrate the codes and categories that I derived the theme 'change of Catbandit's rating styles' from. Furthermore, in the theory building stage, drawing from the data I linked the codes and categories under

this theme all together and made the statements of the relationships about them. In other words, from analysing the codes and categories from Catbandit's attitudes towards assessment, criteria, PD, and his reported assessment practice, I rearranged codes and categories pertaining to his rating style and then drew the connections among them. Figure 7.4 is the outcome of this process. It also shows the categories and codes, concerning Catbandit's rating style, derived from Figures 7.1, 7.2 and 7.3.

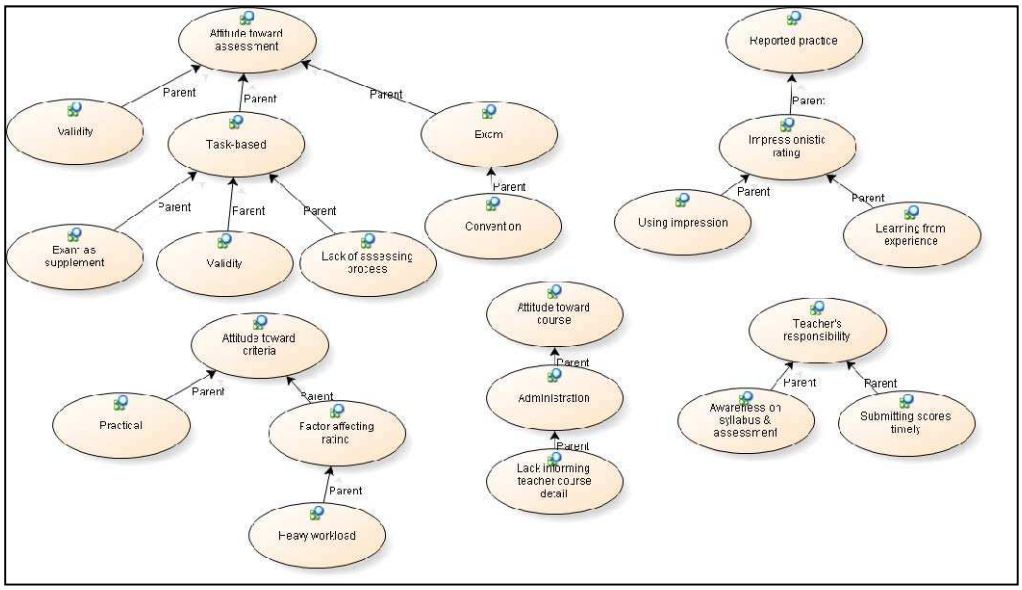


Figure 7.1: The NVivo output model of categories and codes as a map from Catbandit's Interview 1

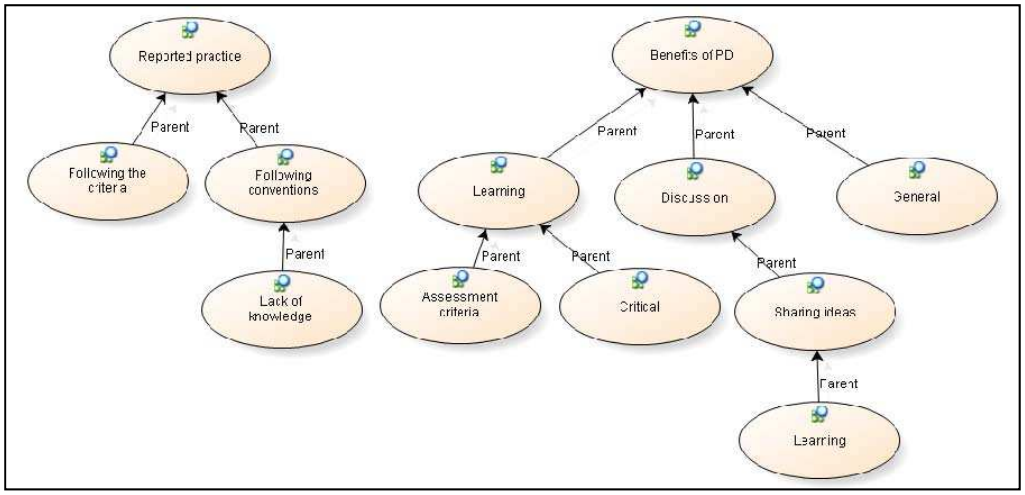


Figure 7.2: The NVivo output model of categories and codes as a map from Catbandit's Interview 2

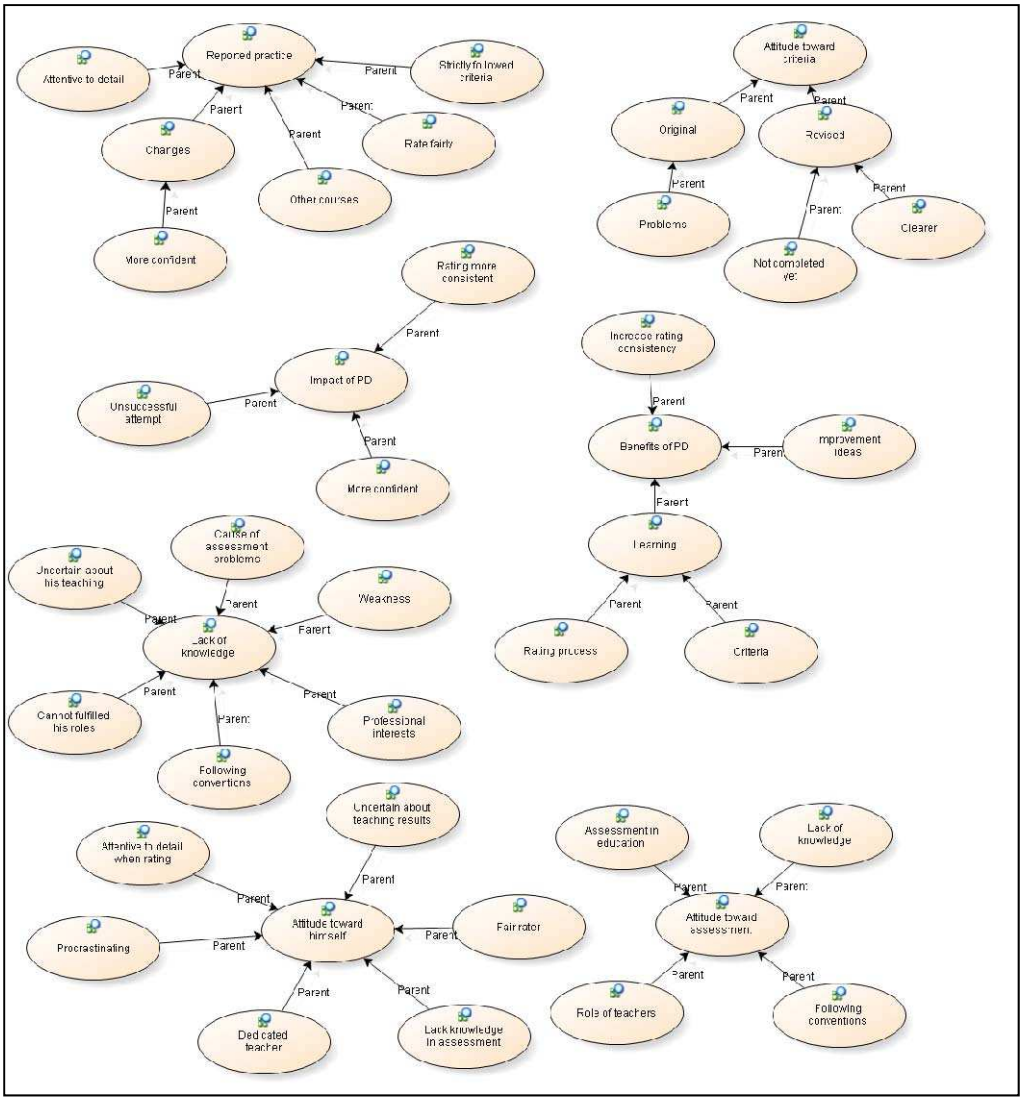


Figure 7.3: The NVivo output model of categories and codes as a map from Catbandit's Interview 3

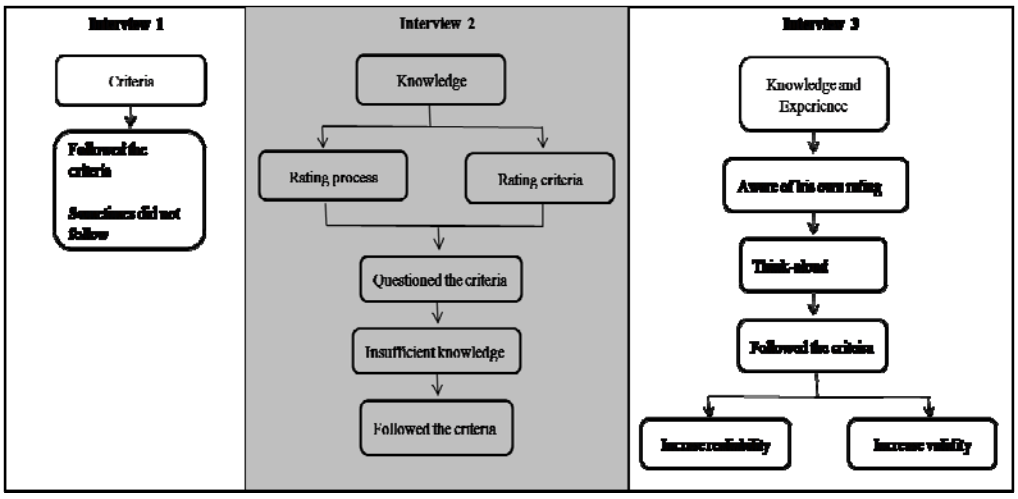


Figure 7.4: A sample of integrated categories

In a further theory building stage or discussion of the changes of each participant, I provide the statements of the meanings drawn from these categories and codes with the references to the raw data (described in the previous chapter). In addition, I revisited the memos and annotations I had taken while coding the transcripts which facilitated the process of integrating categories and theory building. Figure 7.5 is an excerpt from the memos, and Figure 7.6 is a sample of an annotation which helped me refine the categories for the impact of the PD on Catbandit's rating style.

In the first section of this chapter, I explore the themes derived from this analytic procedure for the four participants who changed as a result of participating in the PD workshop. I also provide a figure of which the theme is derived from.

Memo 4 (Interview 1)
24 April 2008
Lack of knowledge in assessment
Catbandit comments on the criteria that he does not understand why they are what they are. Despite that fact, he has been following them. He says that it is like the criteria are there to follow. He points out that he has to follow these criteria because he does not have the knowledge in this field.

Memo 3 (Interview 2)
25 April 2008
The think-aloud: Making rating more consistent
He admits that before the PD, his rating was not consistent. When he was a new teacher, he followed the criteria very strictly. However, the experience has helped him see the patterns of student's works, which changed his rating style – he became less strict (in following the criteria). However, he reports that because for this PD he had to do think-aloud, he had to be aware of the criteria. Thus, it made his rating more consistent. He then explains that he had to do the same (how he did for the think-aloud) to other students' performances – to make the rating consistent.

Figure 7.5: Excerpts from memos on the impact of the PD on Catbandit's rating style

Annotation 6 (Interview 3)
29 April 2008
Lack of knowledge in assessment
It is now very clear that lack of knowledge is a very import factor that influences Catbandit. I think there are many studies on teacher's practices in relation to their knowledge. I have seen many articles about this topic but I did not pay attention to them at that time. So this is another aspect I have to do for the literature review. Perhaps, 'lack of knowledge' will become the higher concept (category) instead of the sub-concept.

Figure 7.6: An excerpt from annotations on the impact of the PD on Catbandit's rating style

7.1 Tracing Teacher Change

Teachers may change in different ways: responding and adapting to the changed conditions and policies, improving their performance and personal growth, and learning through professional activities (Clarke & Hollingsworth, 2002). Furthermore, teacher change does not necessarily mean they do something differently but a change in their awareness (Freeman, 1989). In addition, studies of teacher change have been directly associated with a PD (Burns, 1992; Clarke & Hollingsworth, 2002; Richardson, 1996). Therefore, the main investigation of teacher change in the present study focuses on the changes of teachers as the result from having participated in the PD workshop. Furthermore, changes included in this discussion are the reports of these teachers rating their students differently, and the increase of their awareness in various issues in assessment. In this section, I take turns in exploring the changes of each individual participant based on the process described above and the data presented in the previous chapter.

7.1.1 Participant 1: Catbandit

From analysing the interviews with Catbandit, the main themes that emerged include the impact of the PD on his rating style and his attitudes toward teacher's roles in assessment.

7.1.1.1 Being critical to assessment and changing rating styles

In terms of the impact of the PD on Catbandit's rating style, it appears that the knowledge in assessment he has acquired from participating in the PD workshop has changed how Catbandit views the rating process and, consequently, affected how he rates students' performances. Figure 7.7 below, derived from categories and codes, illustrates this impact.

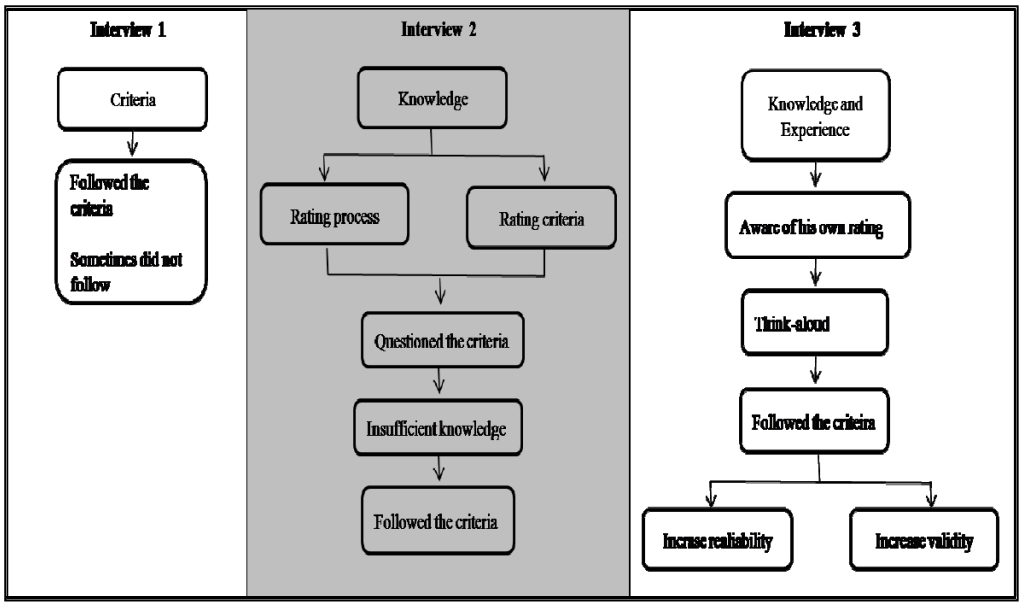


Figure 7.7: Catbandit's rating styles

First of all, drawing from the data, knowledge in assessment allows Catbandit to critique the assessment being used in the Department. According to the interviews, the knowledge in the rating process and rating criteria Catbandit has acquired in the PD allows him to comment on the rating criteria, which he did not do before participating in the PD. The data from the second interview shows that Catbandit previously simply followed the criteria and the conventions in assessment without questioning them. This was because he did not have sufficient knowledge. From the PD, however, he has learned about rating process and rating criteria. In addition, from the experience in rating the samples of students' performances in the PD, he became aware of some problems of the criteria. In other words, the PD has made him become

more critical to the assessment being used. However, the data indicates that he still follows the assessment conventions in the Department because he does not have sufficient knowledge to do anything about it. Furthermore, despite the fact that he is aware of the problems of the criteria, it appears that Catbandit still relies on the criteria as the main guideline because it is the best thing he can do. This may imply that what he has learned from the PD is not enough to allow him to take any roles in solving the problems.

Moreover, grounded in the data, it seems that knowledge of assessment affects Catbandit's rating style. The PD has made him aware that following rating criteria could increase the validity and reliability of ratings. Without knowledge of assessment, Catbandit's rating was not consistent because he sometimes followed the rating criteria and sometimes did not. From the first interview, the data suggests that Catbandit tried to follow the rating criteria as strictly as possible in rating students' performances. However, because of the amount of performances he had to rate, as well as coordinating the course, he sometimes relied on his impression when rating. The data from the second interview supports that this inconsistency in rating is also caused by his insufficient knowledge in assessment. However, the PD may have made Catbandit become more self-consistent when rating students' performances, as indicated in Interview 3. In this interview, it seems that Catbandit has tried to follow the criteria as a result of what he has learned from the PD, especially about the factors affecting rating, and having the experiences in think-aloud and sharing rating experiences with the participants. In contrast to the first and second interviews of which the data indicates that Catbandit used to follow the assessment conventions and the criteria because he did not have sufficient knowledge in assessment, in the third interview, it appears that he has tried to follow the criteria because he wanted to increase the validity and reliability.

7.1.1.2 Realising roles of teachers in assessment

In addition to increasing self-consistency in rating, and allowing Catbandit to critique on the assessment, the PD has also increased his awareness of teachers' roles in assessment. Figure 7.8 shows the categories and codes that this theme emerged from.

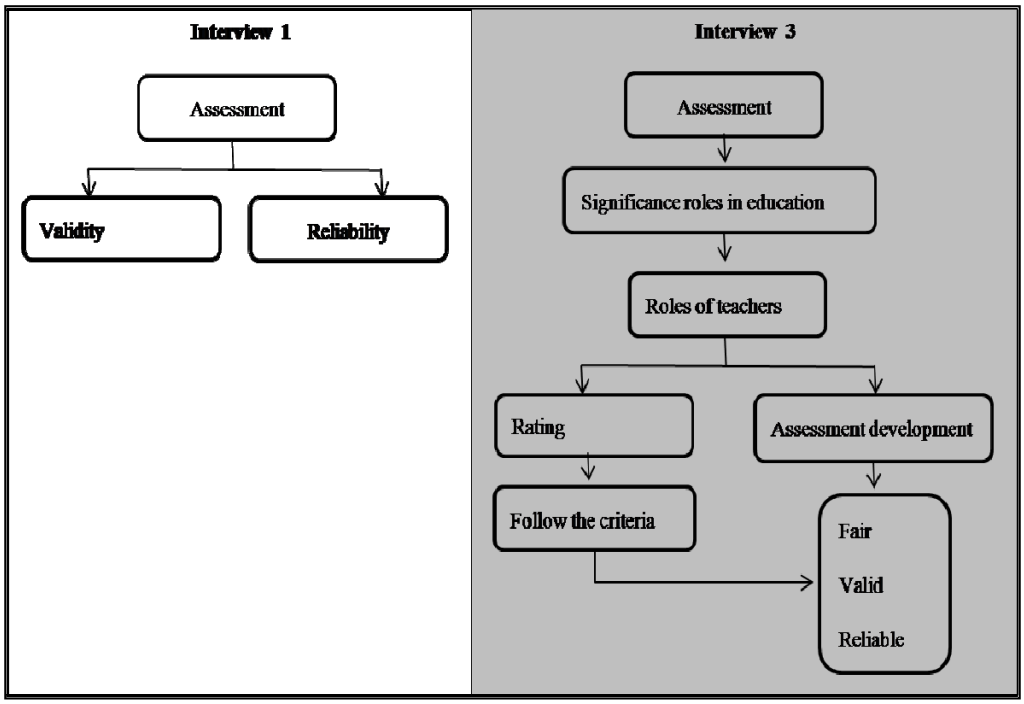


Figure 7.8: Catbandit' attitudes toward roles of teachers in assessment

In the first interview, Catbandit might be aware of the importance of validity and reliability in assessment, but he did not know that teachers have significant roles to play in increasing the quality of assessment. In this interview, it appears that the most crucial aspects of assessment Catbandit was aware of were what to assess and how to assess students, and the reliability of rating. It could be argued that the knowledge in the rating process Catbandit has acquired from his involvement in the activities in the PD workshop has increased his awareness of the roles teachers play in ensuring validity and reliability in assessment. In Interview 3, the data indicates that Catbandit believes that it is the responsibility of teachers to make the assessment valid, reliable and fair. The interview also suggests that Catbandit has realised that because assessment is considered as a very important aspect in education, teachers have two

crucial roles in assessment. The first role is the role of rater. Teachers, who are also raters of their own students in this context, have to follow the rating criteria in order to achieve fair, valid and reliable assessment,. The second role is the role of assessment developer. The teachers, who have to develop and create assessment, must try their best in making the assessment fair, valid and reliable.

7.1.1.3 Summary

The PD is similar to rater training, which helped Catbandit become more self-consistent when rating students' performances. Because in the PD the participants learn about the rating process and rating criteria, Catbandit became aware of the roles of raters and rating criteria; thus, he tends to follow the criteria in order to increase the reliability of their ratings. Whereas a rater training may primarily increase inter- and intra-rater reliability, the PD could also provide Catbandit with knowledge in assessment, which allows him to critique the assessment being used in the system. Furthermore, when Catbandit is provided with knowledge and experience in assessment, he may feel that teachers have a role to play in making sure that assessment is fair, valid and reliable. In addition, as he is aware of the impact of the quality of assessment on students, he would try his best to ensure the quality (i.e. fairness, validity and reliability) of the assessment so it would have positive impact on students.

7.1.2 Participant 2: Papone

From the data overview, it is implied that the PD has given Papone opportunities to critique the past, present and future of his assessment practices. He has become aware of the problems pertaining to the rating criteria and planned to apply the knowledge in assessment he has learned for future applications.

7.1.2.1 Becoming critical of past and present practices

The data from the interviews with Papone implies that the PD has helped him to become more critical to his assessment practice of the past and present. Figure 7.9 illustrates the categories and codes pertaining to Papone becoming more critical of the assessment practice in the Department.

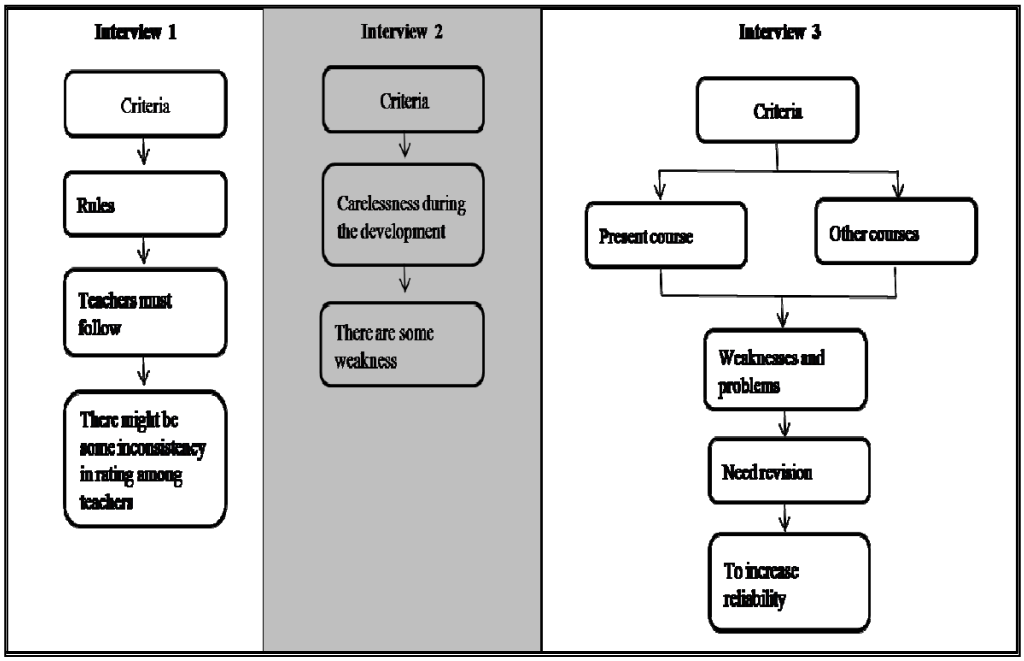


Figure 7.9: Papone's attitudes toward assessment

It is crucial to point out that Papone is the developer of Task 1, FE 2, which was the focus in the PD. He developed both the course materials and assessment for this task. Therefore, he has the information on how the course and its assessment were developed. The data from the interviews with Papone illustrates that the PD activities in which the participants evaluated the assessment are the process of deconstructing the experience of developing the assessment for Papone. The data also implies that these activities have caused him to question his past practice in developing the assessment. Before attending the PD, in Interview 1, Papone was not aware of the problems of the existing rating criteria. He perceived the criteria as rules which teachers have to follow; thus, he followed the criteria very strictly without any question. Although the data suggests that Papone was aware of the inconsistency of

the teachers' ratings, he believed that teachers have to try their best to strictly follow the criteria. However, the activities in the PD workshop have helped Papone to realise the problems pertaining to the quality of the rating criteria he developed. That is Papone began to question the quality of the rating criteria used in the Department. In Interview 2, the data reveals that Papone has realised that the criteria (for Task 1, FE 2, which was used in the PD) reflect some weaknesses of the assessment of the course. In other words, Papone has become aware of the problems and weaknesses of the criteria. The major problem he has discovered is the fact that the criteria do not cover all necessary aspects which should be assessed, but otherwise includes some irrelevant ones, when comparing them against the objectives. More importantly, the PD allows Papone to critically evaluate the practice of the Department in developing the assessment. Because he developed this set of criteria, prior to participating in the PD, he might not be able to see its weaknesses. However, the PD has demonstrated to Papone that there are problems with the criteria, especially on how it was developed. After the participants have shared their opinions towards the criteria and Papone has learned about the process of developing rating criteria, he became aware that the problems pertaining to the criteria were created during the development process. Grounded in the data, the PD has made Papone realise that these problems were caused by the carelessness of the developers, in which he was a part of, during the development process; while he did not realise this before participating in the PD.

Furthermore, the experiences in the PD extended Papone's attention to the present practices in assessment of other tasks in the FE 2 and other courses. Apart from realising the problems of his previous practice in developing the assessment, Papone has also taken notice of the problems of the assessment of other courses. The data from the third interview entails that he has noticed problems associated with assessment in other FE courses, especially their rating criteria which should be revised to increase the reliability of the ratings.

7.1.2.2 Projecting the future applications

Another significant theme grounded in the interviews with Papone is how he plans to apply the knowledge he has acquired from the PD in the future. Apart from being critical of the past and present practice in assessment, the PD also provides Papone with the proper applications in assessment that he will implement in future courses he is responsible for. Figure 7.10 below shows the categories and codes contributing to this aspect of the impact of the PD on Papone.

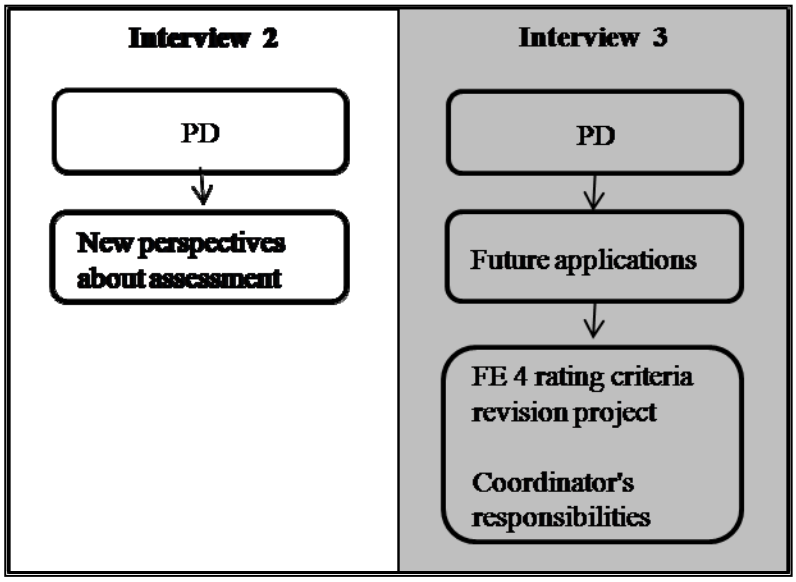


Figure 7.10: Papone's perspectives of future applications

In Interview 2, the data shows that Papone has learned a great deal about assessment from the PD which provides him with new perspectives about the past and present practices in assessment. The data from the third interview also supports that the PD has given Papone the opportunity to learn about assessment which he will be able to apply in the future. It appears that he has volunteered to take part in the revision of the assessment criteria for FE 4 in which he believes that he will be able to use the knowledge he has learned from the PD in revising the criteria for this course. Furthermore, as grounded in the data, Papone, as the coordinator of FE 2, will be able to explain to the teachers about the revised criteria because of what he has learned from participating in the PD. Before the PD, the data indicates that he did not

understand the rationale or principles behind the existing criteria even though he was the person who developed them. After the PD, however, Papone has learned the process of revising the rating criteria. Therefore, he believes that though other teachers have not participated in the PD, he can explain to them the basic principles of how the criteria are developed. The data also indicates that Papone believes that when these teachers understand the underlying principles of the rating criteria, they will follow them.

7.1.2.3 Summary

The PD provides the opportunities for Papone to critically re-examine his past and current practices in assessment. With close examination, he could discover the problems of what he has done and the weaknesses of what he is currently doing. The theoretical and practical knowledge Papone has learned from the PD may also offer him directions in improving these problems and weaknesses. Moreover, he might become aware of how he could make use of this knowledge in the future.

7.1.3 Participant 3: Tanya

Grounded in the data, it seems that the PD provided Tanya with opportunities to learn about her rating style, and, consequently, establish her own way of rating of her students' performances. As a result of this process, Tanya has become confident and self-consistent in her ratings.

7.1.3.1 Deconstructing and establishing rating style

One of the important central themes emerged from the data is that the PD provides the opportunities for Tanya to deconstruct her rating style, and, consequently, establish her approach in rating her students' performances. This theme is illustrated by the figure 7.11 below, which is comprised of the categories and codes derived from the interviews.

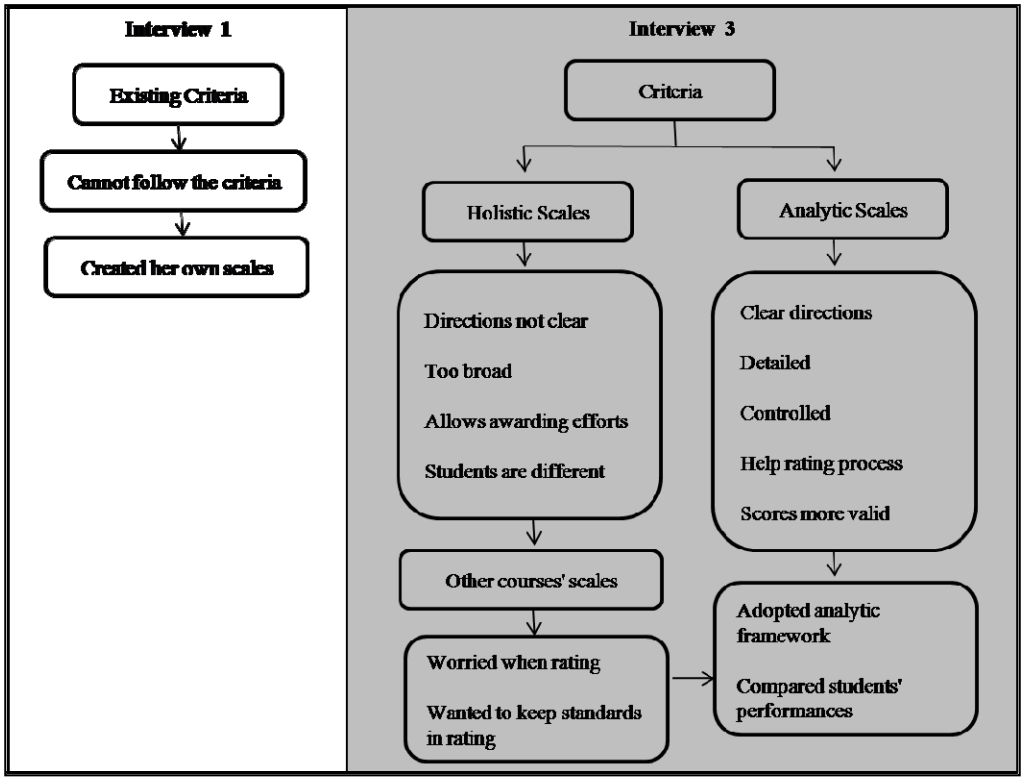


Figure 7.11: Tanya's awareness of problems with assessment

The data from the interviews indicates that the experiences from the PD workshop have helped Tanya understand her rating style and, consequently, develop her consistent rating style. In the first interview, Tanya was working as a part-time teacher with approximately five months teaching experience. She becomes a full-time member before the second interview. According to this interview, which was prior to participation in the PD, the rating experience Tanya has had with ratings has been a confused one because she could not follow the rating criteria and did not understand why. The data indicates that Tanya thought that not having much experience in rating was the reason; in order words, she was not familiar with how to use the criteria to rate students' performances. To solve the problem, she created her own checklists, based on the criteria, using Microsoft Excel, to use when rating students' performances. In addition, according to the first interview, the data shows that Tanya does not have knowledge about different types of rating scales.

However, grounded in Interview 3, the knowledge she has gained from the PD has made her realise that there are two main types of rating scales: holistic and analytic scales, and the scales used in the course are holistic scales. In addition, the PD has made Tanya realise that a holistic rating scale does not fit her rating style because she prefers clear and controlled descriptors of an analytic scale. She has also realised that the checklists she has created as a supplement for the holistic scales are closer to the concept used in an analytic rating scale. Furthermore, the data reveals that when Tanya has become aware of her rating style, she, consequently, has established her own way of rating students' performances. This is indicated by the fact that when she rates students' performances, she creates her own versions of rating criteria based on what she has learned and applied to other tasks in FE 2 and other courses. Moreover, the knowledge about different types of rating scales allows her to assert her belief that holistic criteria are not appropriate to use in this context because they provide clear directions with more detailed and controlled descriptors which help teachers' ratings. The data also suggests that the PD has made her aware that this would help increase the validity of the scores.

7.1.3.2 Becoming confident when rating

Another crucial theme that emerged from the analysis of Tanya's interviews is the impact of the PD on her confidence in making judgements on her students' performances. It is grounded in the data that the experience and knowledge acquired from the PD has made Tanya more confident when she rates the students' performances. Figure 7.12 below illustrates how this theme is derived.

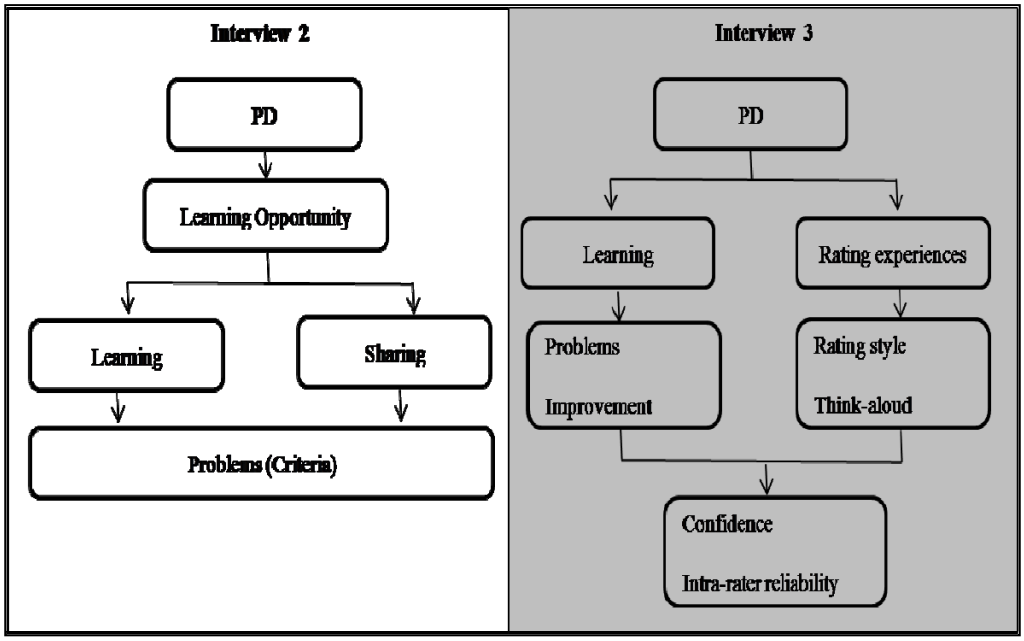


Figure 7.12: Tanya's learning about assessment

As pointed out in the section above, before attending the PD Tanya thought that the rating criteria were appropriate; however, she could not follow them without her spreadsheet checklist. More importantly, she believed that the problem was caused by her limited experience or her personality (as discussed in Section 7.1.3.1 above). However, the PD has illustrated to her that other participants also have had similar problems. The data from Interview 2 shows that Tanya has realised that the rating criteria were the source of the problems she and other participants have had in rating. In addition, in Interview 3, it becomes apparent that Tanya has become aware that it is the criteria that have been the main problem, not her inexperience. In consequence, this realisation has made her feel more confident in how she has been rating her students' performances (i.e. her style as discussed in Section 7.1.3.1). Furthermore, the data implies that when her confidence in making judgements when rating students' performances has increased, Tanya has also become more self-consistent in her ratings. This is supported by the fact that she has been following her rating criteria very strictly, as reported in Interview 3.

7.1.3.3 Summary

The PD offers the opportunities for Tanya to scrutinise her rating styles as well as share them with other participants. In understanding her rating style, and learning about rating process in the PD, Tanya could establish her consistent rating style. Consequently, she could become more self-consistent in rating her students' performances. In other words, the PD serves as rater training for teachers but it might also contribute to a more long-lasting impact.

7.1.4 Participant 4: Wanwisa

The data reveals that Wanwisa gradually changes her attitudes toward the rating criteria, roles of teachers in performance-based assessment, and her behaviour in rating.

7.1.4.1 Recognising possibilities of changes

From analysing Wanwisa's interview data, the central theme emerging from the impact of the PD is her recognition of the possibilities of change in assessment in the Department. Figure 7.13 shows the categories and codes from the interviews representing this recognition. The data implies that the PD has shown Wanwisa that the assessment in the Department can be changed and improved with the influence of those in authority positions.

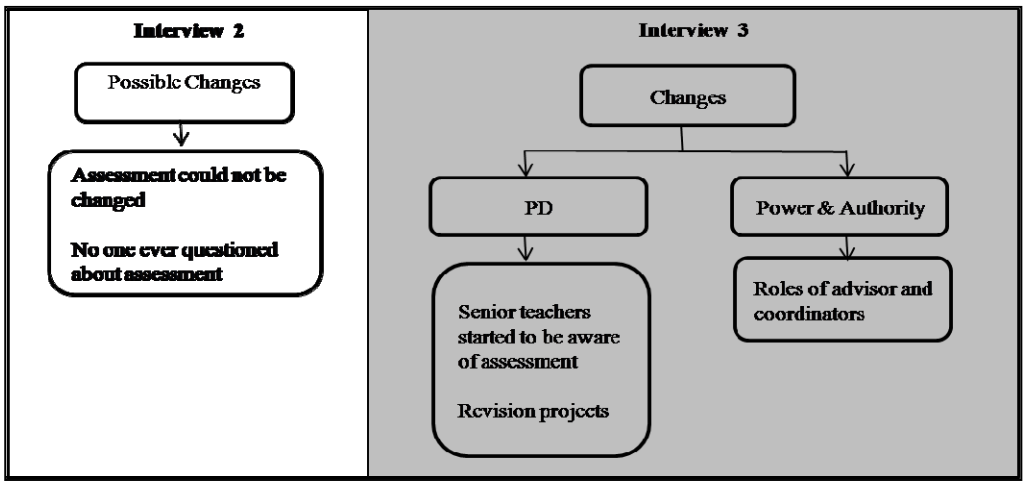


Figure 7.13: Wanwisa's recognition of possible changes in assessment

What the PD has demonstrated to Wanwisa is that changes in assessment are possible. In Interview 2, the data implies that prior to the PD she did not think changes in assessment were possible in the Department because none of the members of staff ever questioned or raised any issues concerning assessment. However, Wanwisa has recently noticed that the teachers, especially those who were in administrative positions such as the Advisor of the FE courses, have begun to discuss about improving assessment in the courses. The possibilities of changes in assessment have been emphasised in the third interview. In this interview, the data indicates that Wanwisa has noticed that because of the PD programme, senior teachers have started to be aware of issues in assessment, as there have been talks about assessment revision projects. Therefore, Wanwisa has been convinced that changes in assessment are possible.

Furthermore, the data reveals that from Wanwisa's point of view, teachers who are in the administrative positions play very important roles in influencing changes. Grounded in Interview 3, teachers in the Department have to follow the coordinators and the coordinators have to follow the Advisor. It is indicated in the interview that although some of the suggestions or rules made by the Advisor were impractical or unreasonable, the coordinators had to follow them. Wanwisa also admitted that as a coordinator herself, she has followed the Advisor even though she did not agree. It is possible that because of this power structure within the Department, Wanwisa believed that the coordinators of the courses and the Advisor are those who can implement changes in assessment. In other words, the success of implementing changes depends on those who have authority. Therefore, in order to implement changes, teachers who have authority must get involved in reform initiatives. In the interviews, teachers who have authority also include those who have expertise in a particular area. The data indicates that the PD has demonstrated to Wanwisa that I (the researcher) have the authority in assessment (because I have been conducting a doctoral research in assessment); therefore, I have the authority to

initiate change in assessment. This is illustrated by the fact that Wanwisa believes that many projects in assessment in the Department were the effects of the PD programme. Therefore, the fact that the Advisor and senior teachers have recently paid attention to assessment (see Section 7.3 below) reinforces Wanwisa's belief that changes in assessment are likely to take place in the Department. The data from the follow-up study supports that changes in assessment actually have taken place after the main study.

7.1.4.2 Realising roles of rating criteria and teachers in rating process

Another important theme that emerged from the data is the impact of PD on raising Wanwisa's awareness of the roles teachers play in assessment. Because Wanwisa has realised that the consistency of teachers in rating affects students' grades and learning, she becomes aware of the role of rating criteria in directing teachers' ratings. This awareness is represented in Figure 7.14 which is comprised of categories and codes contributing to this realisation.

The data from the interviews with Wanwisa shows that the PD has made Wanwisa aware that teachers and rating criteria have the impact on the reliability of the assessment. In the first interview, the data indicates that Wanwisa did not think that teachers' subjectivity and inconsistency in rating affect students' grades. In the second interview, after having participated in the PD, Wanwisa still did not recognise the relationships between teachers' reliability in rating and the roles of rating criteria. This is supported by the fact that she maintained that teachers have to depend on their own judgements when rating students' performances.

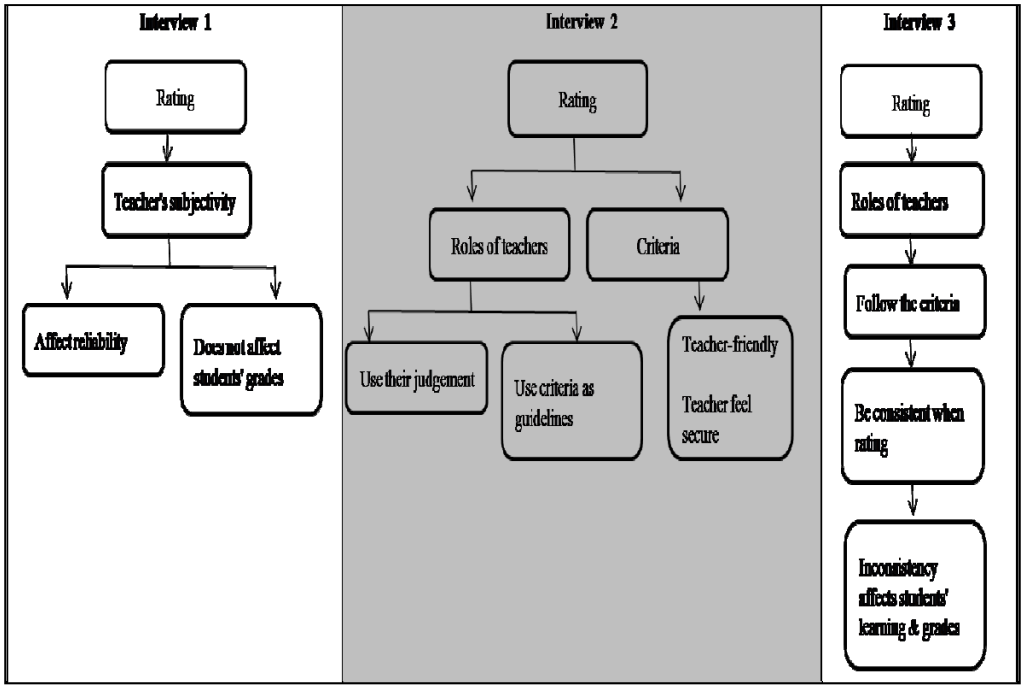


Figure 7.14: Wanwisa's attitudes toward rating criteria in rating process

For Wanwisa, the rating criteria were only guidelines and teachers should not allow the criteria to dominate their ratings. It can be said that the PD has only made Wanwisa aware the roles of the criteria play in teachers' rating process as she points out that the criteria have to be "*teacher-friendly*" and teachers have to feel "*secure*" when they employ the criteria. However, the data from Interview 3 shows that Wanwisa has started to think differently. It appears that the PD has made Wanwisa becomes aware of the role of rating criteria in increasing teachers' reliability in rating students' performances. Grounded in this interview, Wanwisa believed that teachers have to follow the rating criteria, which is different from the previous interviews when she believed that teachers should only use the criteria as their guidelines. This implies that Wanwisa has learned from the PD that it is the responsibility of teachers to make rating consistent, which can be done by following the criteria.

Furthermore, the PD has illustrated to Wanwisa that the inconsistency in rating has a direct impact on students' learning as well as their final grades, which has

also affected how she has been using the rating criteria. Figure 7.15 below illustrates the change of Wanwisa's use of the rating criteria.

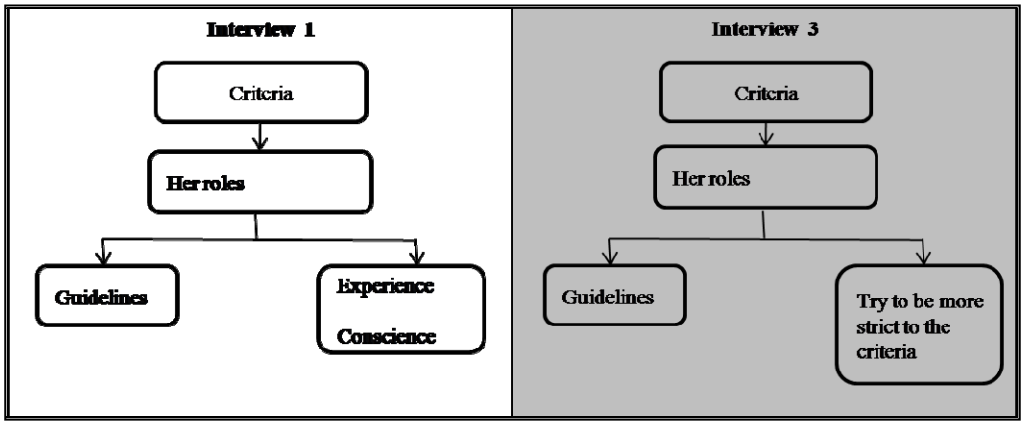


Figure 7.15: Wanwisa reported practices in rating criteria

The data from Interview 1 implies that at the beginning of the PD, Wanwisa perceived the rating criteria as guidelines. When rating students' performances, she relied heavily on her experiences and conscience. That is she used her general impression when rating instead of following the criteria. However, after having participated in the PD and realising the impact teacher's rating behaviours have on students, it appears that Wanwisa has changed her rating styles. Although she still used the criteria as guidelines, the data indicates that she has tried to follow the criteria as much as possible as she has become aware that teachers have to follow the criteria to make rating reliable. This also implies that changes in one's behaviour may take longer than changes in one's belief.

7.1.4.3 Summary

The knowledge and experiences acquired in the PD could raise Wanwisa's awareness of the significant roles of rating criteria, and raters, in the rating process of performance-based assessment. This awareness, in consequence, leads her to attempt to strictly follow the criteria in order to increase the reliability. Furthermore, Wanwisa also realises the significant roles teachers play on assessment and the impact of their practices on students.

7.1.5 Participant 5: Songsri

Songsri maintains throughout the study that she has not changed, or does not need to change, because she believes that her beliefs and practices in assessment are already in line with the principles advocated in the PD workshop. Songsri's resistance to change will be explored in Section 7.2.3.

7.1.6 Summary

It can be summarised that the PD is like rater training, which helps teachers become more self-consistent when rating students' performances. The PD also provides the participants with knowledge and experiences in assessment which allow them to critique the assessment being used in the system. They become aware of the significant roles of rating criteria and raters in the rating process of performance-based assessment. Consequently they realise the significant roles teachers play on assessment and the impact of their practices on students. In the following section, I offer my interpretation of the impact of the PD on these teachers based on the changes described in this section.

7.2 Impact of the Professional Development Programme

It has been argued that assessment has been used as a 'lever for change'. The changes, as the effect of the assessment on teaching and learning, can be either negative or positive. The studies of this kind of change, or known as impact/washback study, have been carried out by language testers in the past decades. In these studies, the changes are the result of the changes of assessment mandated by policy makers, for example, the new national test in Sri Lanka (Wall and Alderson, 1993; Wall, 2005), ASL and EFL tests in Israel (Shohamy et al., 1996), and HKCEF in Hong Kong (Cheng, 2005). According to Rea-Dickins and Scott (2007b), the majority of empirical studies in washback and impact of language testing and

assessment have focused on the investigation of two major international language proficiency examinations, namely the TOEFL (for example, Alderson & Hamp-Lyons, 1996; Wall & Horák 2006, 2008), and IELTS (for example, Hawkey, 2006; Saville & Hawkey, 2004). These changes, very often, are not initiated by the teachers nor do the teachers have any involvement in the introduction or development of the assessment, although the assessment have directly or indirectly caused the impact on the teachers.

In performance-based assessment, in particular, effects of the assessment could take place as a result of the final test product as well as during the test development stages. However, there have not been extensive studies carried out in this area (Turner, 2001). From Turner's (ibid.) observations from empirically derived rating scales in high-stakes performance testing, she advocates for a research into the impact of performance-based assessment and rating scales development on educational settings especially the impact on participants during the construction, validation and implementation of the rating scales. Although studies into the efficiency of empirically derived scale have been conducted (e.g. Knoch 2007a; 2007b), to date, there has not been any study investigating the impact of such a scale on the development team who are teachers and who actually use the scale. The present study, therefore, aimed to shed some lights on the impact of taking part in the development of the empirically derived scale integrated in the PD workshop on teachers who participated.

From the analysis of the data presented in the above section, the PD has a positive impact on the rating styles of the participants, and, increased their awareness in the roles they can take in contributing to positive changes in assessment practice. This revelation reinforces Hamp-Lyons' (2007b) stance that a PD, a teacher empowerment process, can lead to positive change in teachers. In the following sections, I explore the impact of the PD in assessment on teachers who participated.

7.2.1 Increasing rater reliability

According to Hamp-Lyons (2007b), a rater training is part of professional development for teachers. She asserts that teacher raters will become more self-consistent in the assessment in their classroom when they receive adequate preparation. Hamp-Lyons (ibid, p. 499) stresses that:

> Since teachers are both interlocutors and raters in their own classroom, professional development can capitalize on the variability of response to language performances and help teachers to, first, deconstruct their own preferred ways of responding to student's language, and then to establish a consistent approach to responding to student work.

The findings from the present study support this. The findings reveal that four teachers (out of five) who participated in the PD workshop have changed their rating styles and become more self-consistent in ratings.

7.2.1.1 Changing rating styles

The knowledge and experiences the participants have acquired from the PD have had the impact on their attitudes towards the criteria and how they later apply them. The interviews with Catbandit revealed that from participating in the PD, he has chosen to follow the rating criteria when he rates student's performances. When he first started his teaching career, he followed the rating criteria very strictly, but he gradually relied more and more on his impressions as he gained more rating experience. The data indicates that this is because he did not have an understanding about the rating criteria: how they were developed and why they were developed this way. However, after having participated in the PD, in the final interview in the main study, Catbandit has been following the criteria very strictly because he believes that it helps increase the reliability of the ratings. The data implies that Catbandit has now realised that

teachers have to follow the criteria in order to make certain the ratings are reliable across different sections of the same course.

Similarly to Catbandit, the PD has also made Wanwisa become aware of the significant impact of the rating criteria on the reliability of ratings. Consequently, she has tried to follow the criteria when rating students' performances. Grounded in the first two interviews, Wanwisa used to believe that the rating criteria were only guidelines and teachers had to rely on their personal judgement about students' abilities when rating their performances. However, the data from the third interview shows that Wanwisa has been thinking differently. It appears that she has acknowledged that following the rating criteria increases the reliability of rating and become aware that teachers must follow the criteria. The data also indicates that Wanwisa has been trying to do the same; that is following the criteria when rating. In summary, after having participated in the PD, Catbandit and Wanwisa have changed their rating styles by trying to follow the rating criteria because they believe that this can increase rater reliability.

7.2.1.2 Being more self-consistent

The PD provides the participants opportunities to examine their own rating styles as well as compare them with others. With the knowledge and experiences gained from the PD, the participants consistently apply their rating styles to other courses. According to the interviews with Tanya, it is grounded in the data that the PD allows her to understand her own rating style from discussing issues in the rating process, especially from scrutinising a sample of her think-aloud protocol. In addition, the experiences of studying the samples of other participants' think-aloud protocols have built her awareness of other participants' rating styles. These experiences have helped her create her own versions of rating criteria and she has been using these rating criteria when rating students' performances. Furthermore, the PD has helped increased her confidence with her rating style as grounded in the data in Interview 3

that Tanya has the confidence when making judgements about her students' performances. The data also reveals that she also has applied the same rating style to other tasks in FE 2 (the focus of the PD workshop) and other FE courses.

In addition to Tanya, the data from the interviews demonstrates that Catbandit and Wanwisa have become increasingly self-consistent in following the criteria (as previously discussed in Section 7.2.1.1 above). Nevertheless, differently from Tanya, they did not create their own versions of the rating criteria. Catbandit and Wanwisa changed their rating styles from not following the criteria to following the criteria. To summarise, the rating styles of Tanya, Catbandit and Wanwisa have become more consistent after having participated in the PD.

7.2.2 Being critical to assessment

Not only participating in a PD allows teachers to deconstruct their rating styles and establish a consisting rating, it also encourages teachers to critically evaluate the assessment as well as raises their awareness in the roles they can play in influencing assessment practices in the Department. Hamp-Lyons (ibid., p. 492) stresses that:

> Pre-service and in-service professional preparation and development for teachers provides an essential opportunity for teachers to critique their position in the education society, identify points of opportunity and mechanisms to influence education planning, including assessment, and to find ways to contribute to positive change.

The data from the main study and the follow-up study reveals that four participants have become critical to the assessment. They recognise the problems in assessment, develop the awareness of teacher's roles in assessment, and become critical to the changes in assessment practices in the Department.

7.2.2.1 Recognising problems

The activities in which the participants shared their opinions about the assessment tasks and the rating criteria, as well as their experiences in rating students' performances, make the participants more critical of their assessment practice. Grounded in the interviews with Papone, the PD has made him aware of his past mistakes in developing the rating criteria, as well as the current problems pertaining to the assessment tasks and their criteria. Prior to participating in the PD, Papone did not recognise the problems with the rating criteria, though he is aware of the problem of rater reliability. However, after he had taken part in the PD, Papone has begun to be critical of the assessment practice in the Department. The data confirms that he has become aware of the problems relating to assessment of the course that he developed, and these problems were partly caused by the lack of knowledge and carelessness of the material writing team during the developing process. It is included in the data that prior to the PD, Papone did not know that it is important to match the objectives of the assessment tasks and the rating criteria, which he learned from the PD. Moreover, Papone has questioned the rating criteria of other courses. He believed that they also need to be revised.

Similarly, Tanya has also recognised the problems with the rating criteria. Before she participated in the PD, the data reveals that Tanya believed that she could not follow the criteria because of her limited rating experiences. However, from her experiences in the PD, she has become aware that it is the criteria that caused her confusions when rating her students' performances. Furthermore, Catbandit has also become critical of the rating criteria. The interviews indicate that after having participated in the PD, Catbandit has recognised problems with the criteria of the courses; thus, he believes that these criteria need to be revised to increase the rater reliability. In summary, the PD has provided Papone, Tanya and Catbandit necessary

knowledge and experiences in performance-based assessment to be able to critically evaluate the assessment, especially the rating criteria used in the Department.

7.2.2.2 Being aware of teacher's roles in assessment

In addition to identifying the problems in assessment, the PD also provides the participants the opportunity to critique their personal roles, and other teachers' roles, in influencing changes in assessment within their education society. From the analysis of Wanwisa's interviews, the data shows that the PD allows her to realise the various roles different teachers in the Department have to perform in order to improve the assessment. According to the interviews, Wanwisa has recognised the effects PD has had on the Department due to the fact that the senior teachers have begun to pay more attention to assessment. Therefore, she is convinced that the teachers with expertise in their areas should take part in initiating improvements in the Department. Furthermore, the PD has demonstrated to Wanwisa that in order for any change to be successfully implemented, it is important that senior teachers, who are in administrative positions, take the lead because of the seniority culture in the Department (see the discussion on the assessment practice of the department in Section 7.3 below). As for teachers in general, Wanwisa has realised that one of their roles in assessment is to confirm the reliability of their ratings. As described in Section 7.1.4.2, Wanwisa has begun to believe that following the criteria can increase rater reliability, the data also confirms that she has been trying to follow the criteria, though previously she used the criteria only as guidelines. In addition, she advocates that other teachers have to strictly follow the criteria. Furthermore, grounded in the interviews in the follow-up study, Papone also agrees that the PD has initiated the changes in assessment in the Department.

In the same vein, the PD has illustrated to Catbandit the roles teachers play in rating their students' performances include the roles of assessor and assessment developer. According to the interviews, the PD has made Catbandit aware that

teachers, as assessors, have to follow the rating criteria when rating the students' performances in order to make assessment fair, valid and reliable. In addition, the PD has shown him that because assessment has a very high stake in educational system, therefore, when developing assessments teachers, as assessment developers, have to make sure that they are fair, valid and reliable. In conclusion, the PD has made Wanwisa and Catbandit aware of their own roles, and other teachers' roles, in improving the quality of assessment in the Department.

7.2.2.3 Being critical to changes in assessment practice

Finally, the experiences and knowledge acquired from PD has helped the participants to recognise the changes in assessment that have taken place in the course of the present study (for a discussion on these changes, see Section 7.3 below). Consequently, they have been able to critically evaluate them. The data from the interviews in the main study shows that Wanwisa has been aware of initial reforms in assessment which are taking place in the Department, which she has recognised from the fact that senior teachers have begun to raise issues in assessment in the meetings and there have been discussions about improving assessment, especially the rating criteria. According to the data from the follow-up study, the department took a major step in reforming assessment by revising the rating criteria for all tasks for FE 1 and FE 2. Grounded in the interviews, the data suggests that Wanwisa was satisfied with the changes. However, she proposes that the Department should not implement any more changes, but instead focus on, and improve upon, what has already been changed.

On the other hand, the data from the follow-up study indicates that Catbandit and Papone, who were part of the team in revising the rating criteria for FE 1 and FE 2, did not agree with the Department's reform initiatives. The data shows that though they agreed with the revision of the rating criteria, they did not agree with the process the Department took in the revision. Because Catbandit and Papone have learned

about the principles underlining empirically derived rating scales (see Section 2.3.2) from the PD, they were very critical when the Department made the decision of change based on the intuitions of teachers. According to the interviews with Catbandit, the changes (i.e. in rating criteria) should be based on empirical studies or assessment principles. Thus he did not agree when the Department decided to revise the criteria based on teachers' intuitions. Similarly, according to Papone, in order to implement change, the team has to begin by questioning why they need to change and investigating the problems and the causes of the problems. In other words, the Department needs to conduct an empirical study before implementing the reform, similar to the process in revising the rating criteria in the PD workshop. To summarise, the experiences and knowledge from the PD workshop allow Wanwisa, Catbandit and Papone to critically evaluate the reforms in assessment that take place in the Department.

7.2.3 Resistance to change

Different from the four participants (Catbandit, Papone, Tanya, and Wanwisa) discussed above, Songsri maintains throughout the study that she has not changed or does not need to change because she believes that her beliefs and practices in assessment are already in line with the principles advocated in the PD workshop. This indicates that Songsri remains unchanged by the reform input. In explaining Songsri's resistance to change, I apply the self discrepancy theory to investigate the factors behind her resistance.

The absence of teacher change as a result of reforming initiatives could be explained by the role of dissonance as adopted in Kubanyiova's (2009) study. In her study, Kubanyiova has adopted the concept of ideal selves and, specifically, self discrepancy theory (Higgins, 1987, 1996; cited in Kubanyiova, 2009) to explain the absence of teacher change. The findings of Kubanyiova's (in press) study have shown that discrepancy between actual and ideal selves is critical in teacher change. It seems

that if the teachers are aware of the limitations of their current assessment practices, and at the same time aspires to improve them, they are more likely to engage with the PD input at a deeper level, a key condition for conceptual change. However, Kubanyiova (2009, p. 328) stresses that:

> without individuals' awareness of a discrepancy between their actual and possible selves, which is accompanied by dissonance emotions, there is no gap to be reduced and therefore no motivation to further engage with the reform input.

In other words, when teachers are not aware of their limitations of the assessment practices, they do not seek to change. Grounded in the data, Songsri has proven to be the case.

First of all, the data suggests that Songsri's attitude toward the assessment has not changed at all. It appears that Songsri has recognised the weaknesses of the assessment from the beginning (i.e. in the first interview). However, she cannot do anything about it because she does not have any authority to initiate any change. The data from the second and third interviews also confirms that her attitude has not changed. In these interviews, Songsri maintains that the assessment being used in the Department is not appropriate but she does not have any power to intervene. In the same vein, the PD has not had any impact on Songsri's practices in assessment. The data from the first interview indicates that Songsri needs to count the lexical items when she rates students' written performances. In the third interview, the data reinforces that she will continue counting despite the fact that the participants in the PD have decided not include counting lexical items in the rating criteria. In addition, the data from the follow-up study confirms that Songsri has not changed.

Finally, concerning the PD workshop, the data from the interviews indicates that according to Songsri's point of view, the PD workshop is for the other participants to learn, not for her because the PD has only confirmed what she has

already known and what she has learned is not "*innovative*". The data from the follow-up study also confirms that Songsri has not been affected by the PD as the data reveals that Songsri reported that the PD has not had any impact on her rating style (i.e. she still believes that counting the lexical items is the best way to encourage students to learn). Nonetheless, in this phase of the study, it appears that Songsri has recognised that the PD has helped her understand the rating criteria; therefore, she has found that it has become easier to follow the rating criteria when rating students' performances.

It seems that Songsri believes that her knowledge and skills in assessment are already aligned with the core principles advocated in the PD workshop, therefore, she does not find it necessary to engage with the PD input other than on the surface level (i.e. positive appraisal of the PD). This may explain why no obvious change in her assessment beliefs has been traced in this study. Furthermore, Songsri strongly resembles the case of Silvia discussed in Kubanyiova (in press). Because Silvia is satisfied with her instructional practices, therefore, she does not perceive herself having any impact from the reform attempt. Her practices still heavily rely on her prior beliefs and theories.

7.2.4 Summary

To summarise, the PD has created positive changes in four teachers who participated in the workshops. These teachers have changed their rating styles to become more self-consistent. They have also become critical to the assessment as they have recognised the problems in the assessment being used, become aware of the roles of teachers in influencing assessment practice, and been critical to the assessment reforms. In other words, the PD has empowered the teachers in assessment. The reason that one participant has not changed might be due to her lack of awareness of her limitations of the assessment practices.

7.3 Assessment practices in the Department: Preliminary Investigation

From the discussion of the findings presented in the above sections, it has become apparent that the assessment practices in the Department have directly influenced the attitudes and practices of the PD participants. Therefore, in this section, I will briefly reflect on some of the prominent issues of the Department's assessment practices to provide the context for the impact of the PD on the participants in the discussion above. It should be noted, however, that because of the time limitation of a PhD research and the amount of time required, especially in analysing the qualitative data and writing-up the thesis, I have not been able to explore the rich data collected from the field observations pertaining to the assessment practices in the Department. Nonetheless, I include the assessment practices of the Department in this section, particularly the changes referred to by the participants, because I believe that they could clarify the matters previously discussed in this chapter. The topics discussed in this section include the Department's increasing attention to assessment and the attitudes of teachers, who did not participate in the PD workshop, toward the PD.

In the Introduction Chapter, Section 1.1, I point out that prior to the present study, the issues of assessment have been a main problem within the Department but there had not been any substantial or effective attempts to solve these problems, which was the inception of this study. Furthermore, the findings from the pilot study (see Section 4.2.4) reflect the problems of raters and rating criteria in the Department. The results from the investigation of rater behaviours conducted before the main study (see Section5.2.1) also confirm these problems. From my preliminary analysis of the field observations, the Department has started to pay more attention to the problems in assessment since the beginning of the study.

The first attempt the Department tried to improve the assessment was to provide the teachers in the department with knowledge in assessment. The

Department was very fortunate to have Professor Liz Hamp-Lyons, my doctoral supervisor, visit and give a one day workshop (on 26 October 2007) on basic principles of performance-based assessment, particularly on rating criteria and scales. The first half of the workshop was on the principles of performance-based assessment, and the last part was on assessment criteria and scales. The workshop also introduced teachers to 'good assessment practices'. Moreover, at the end of the workshop, Professor Hamp-Lyons suggested the potential benefits of my research to the department.

After this workshop, I have observed that changes in assessment have started to take place in the Department. The changes are pointed out by Wanwisa (as discussed in Section 7.1.4.1 and 7.2.2.3) as well as by Catbandit and Papone (also in Section 7.2.2.3).These changes include the revision of the rating criteria for the FE courses and the introduction of standardisation meetings for teachers. Prior to the present study (i.e. before the pilot study), the department did not hold any standardisation meeting or rater training for teachers. However, the Department provided four standardisation meetings for teachers who taught the FE courses during the course of this study (i.e. from the beginning of the pilot study to the follow-up study). Another change relating to assessment I have observed is the revision of the rating criteria for the FE courses. After I reported the outcome of the PD workshop (i.e. the revised rating criteria for the written task for Task 1 FE 2) and its potential benefits to the Department, the committee decided to revise the rating criteria of the FE course. The participants in the PD workshop, Papone and Tanya, were the key teachers in the revision of FE 1 and 2 rating criteria. Catbandit was also involved in giving comments on the criteria in the revision process. This revision took place after the main study and before the follow-up study.

Another important aspect of assessment in the Department pertaining to the present study is the attitudes of teachers, who did not participate in the PD workshop,

toward the PD. The most influential teachers in the FE courses are the Advisor, who is the leader of the FE courses, and the course coordinators, as reported by Wanwisa (e.g. in Section 6.2.4.3) and Songsri (e.g. in Section 6.2.5.1). The Advisor, who is usually the most senior teacher in the FE team, overlooks the management as well as the academic related issues of the course, such as the developing the courses' syllabus and assessment. For example, the examinations have to be approved by the Advisor before they can be administered. According to my observations, the Advisor recognises the potential benefits of the PD workshop (as she learned from Professor Hamp-Lyons' suggestion in the workshop). While I was at the Department conducting the PD workshop and data collection, the Advisor constantly asked me for advice on preparation for the standardisation meetings and the rating criteria revision projects.

However, I observed one teacher who did not agree with the core principles advocated by the PD. The recently appointed FE 1 coordinator disagreed with the idea of revising the criteria. In one of the Department meetings when I reported the revised criteria from the PD workshop and its potential benefits, and suggested that the criteria used in the FE courses needed to be revised, the FE 1 coordinator resisted introducing changes of the assessment to the course. He argued that the solution to the assessment problems of the course was to improve the management system. In the meeting, he presented the ideas he and the assistant coordinators had planned to implement in the following semester. When I argued that these ideas could only solve the management problems, and emphasised that what was needed was the improvement of the assessment quality by revising the rating criteria, the coordinator was very angry. However, the Advisor stressed that the revision was one of the options which was worth trying. She pointed out that they should try out this new idea and decide later to adopt or reject it. Thus, they decided to revise the rating criteria used in these courses.

To summarise, in this section, I have provided the contextual information concerning assessment practices in the Department since the beginning of the present study; including the revision of the rating criteria and the introduction of standardisation meetings. I also offered my observations of the attitudes of the teachers, who did not participate in the PD workshop, toward the PD. However, it is important to stress here that a more rigorous investigation of the data from the field notes concerning the issues discussed in the section is needed to make any claims that these changes were caused by the PD.

7.4 Conclusion

The PD has had positive impact on the teachers who participated in the PD workshop. The data reveals that the PD has various impacts on the teachers, except one participant who reported not having any impact from the PD. The changes presented include changing in rating style, realising roles of teachers in assessment, becoming critical to assessment practices, deconstructing and establishing rating style, becoming confident in rating, recognising possibilities of change and realising roles of rating criteria and teachers in rating process. Moreover, the follow-up study also confirms the impact of the PD on these teachers.

8 Conclusion, Implications and Limitations

This study investigated the development of five EFL teachers in Thailand who participated in a PD programme in language testing and assessment. The study was conducted in three phases: pilot study, main study and follow-up study. The PD, implemented in the main study, focused on providing the participants with theoretical and practical issues in performance-based assessment. The main focus of the PD was on developing empirically derived rating scales. In terms of research methodology, the study employed a qualitative inquiry approach with the use of interviews, focus groups, observations, and think-aloud. The analysis of the data was guided by Grounded Theory.

The findings from the pilot study, and the preparation stage of the main study, indicated that the problems pertaining to assessment in this research context were the differences among teachers in their views toward the rating criteria and how they applied them, as well as their lack of sufficient knowledge in performance-based assessment. Therefore, a PD programme in language testing and assessment on performance-based assessment, with the focus on rating process, was implemented in the main study to provide the participants with theoretical and practical principles of performance-based assessment. The data from the main study revealed that the PD had a positive impact on the teachers who participated in the PD. Grounded in the data, it is apparent that the participants in the PD workshop have become aware of the problems of the assessments being used in the department, learned about performance-based assessment, realised the roles of teachers play in assessment, and the impact of their rating behaviours have on the assessment. Moreover, they have changed the ways they rate their students' performances in order to increase the consistency by attempting to follow the rating criteria. In other words, the PD has

helped the participants create more intra-rater reliability as well as become critical to the assessment.

Although more research is needed to support the claims I have made in the thesis, having worked with the teachers in the PD workshop and analysed the data, I have found that the present study has generated several broad issues with regard to the implementation of a PD, empirically derived indigenous rating criteria, and collaborative action research.

8.1 Implications for Professional Development Programmes

One of the answers to Malone's (2008) question of what could be done to support and train teachers when they assess the students, I believe, is involving teachers in a PD. The positive comments from the participants concerning the content and how the PD was implemented confirm the recommendation of Brindley (2001), Malone (2008) and Stiggins and Conklin (1993) that the PD has to match what teachers do and already know in assessment. I believe also that the PD has to match teachers' needs in specific context, in addition to Hamp-Lyons' (2003) proposition that teachers need to know how assessment works and what it can and cannot do. Therefore, a rigorous background study of teachers' needs in that particular context is a compulsory step before any implementation of a PD for a PD to have positive impact. Concerning the format of the PD, it is crucial to provide the participants with hands-on experience. Thus, a series of workshops should be provided as an ongoing in-service training for teachers. In addition, the workshops do not need to include many participants, but they should come from different backgrounds - for example, teaching experience and education backgrounds. The workshops, furthermore, should be conducted in an informal manner which allows the participants to be at ease in sharing their opinions and experiences.

From my experience in implementing a PD in assessment for five teachers, I have found that a PD workshop should have the following characteristics:

- *Teacher-centred*: teachers are key persons in deciding the directions of the PD activities
- *Discussion-oriented*: the activities encourage the participants to share their ideas and experiences; the leader plays a minimal role in the discussions
- *Empirical-based*: the participants have opportunities to have hands-on experiences
- *Indigenous*: the activities and discussions are based on the local needs with minimal intervention of external materials
- *On-going*: a PD is done as a series of workshops or on-going in-service training

8.2 Implications for Empirically Derived Indigenous Rating Criteria

The rating criteria developed in the PD was empirically derived from students' performances based on the local context: local purposes of the assessment, local syllabus and the local students. Thus, they can be called 'Empirically Derived Indigenous (EDI) rating criteria' (for the discussion on indigenous criteria, see Section 2.3.2). Though the present study does not claim that EDI rating criteria could create positive washback on teaching, the findings indicate that the EDI criteria development approach has a positive impact on teachers who are involved in the development process.

Grounded in the data, the participants (except Songsri who was resistant to change) in the PD workshop stressed that after having participated in the workshop, they had tried to follow the criteria when rating their students' performances. In other words, these teachers feel obliged to follow the criteria because of their involvement in the EDI criteria development process. They may feel ownership of the criteria

since they spent many hours in developing the criteria. Thus, they try their best to follow the criteria. Moreover, when these teachers interpret the descriptors and use the criteria to rate the students' performances, they tend to be rating the same constructs. This is because they have been familiarised with the descriptors during the development process, as evaluating the descriptors and students' performances is part of the EDI criteria development process. In other words, the EDI rating criteria could contribute to more valid and reliable rating. Nevertheless, more empirical studies are needed to support this claim.

Drawing from the present study, the findings imply that the components of a EDI rating criteria development process should include teachers, course syllabus, samples of students' performances, and an assessment expert. This process involves:

- *Teachers*, who use the criteria, develop the criteria base;
- *Course syllabus*, which also include analysing the course objectives and the objectives of the assessment;
- *Students' performances*, which are from the assessment of that course and will be rated using the developed criteria; and
- *Assessment expert*, who ideally, should be one of the teachers in that context to give professional advices.

Furthermore, drawing from my experience in introducing new practice in assessment (i.e. an innovation), I have observed the followings which might be useful for a future implementation of a PD programme in general:

- It is advisable not to be too ambitious about the number of teachers enthusiastic to participate in a PD programme. It is more important to have participants from different background, such as educational background and experiences.
- Though a PD programme does not have to include many participants, a small number of participants could have tremendous impact on the social system when the programme includes participants who manage the courses (e.g. coordinators

and advisors) and classroom teachers. Thus, in order the implement of innovation to be successful, it is important to involve both groups of individual as early as the stages of innovation development or the initiation of the innovation.

- Because the participants could have different levels of understanding of the innovation being introduced in a PD programme, it is important to provide them with adequate knowledge of that innovation and how it functions.
- The educators or researchers should get involved in all stages of the diffusion of innovation.
- Policy makers or administrators have to be well informed about the potential benefits of the innovation if they are to be able to support and reinforce the diffusion of the innovation. Moreover, they should be made aware that change is a long and complex process, thus, they should not expect a significant change in a short period of time.
- After members of the social system have made a decision to adopt an innovation, they need constant support from other members of the system to integrate the innovation into an ongoing practice of the system. Thus, ongoing PD could be implemented to reinforce their decision, which could contribute to positive changes.

8.3 Implications for Collaborative Action Research: a Reflection

Reflecting on the process of working with colleagues on professional development at an attempt towards positive change, I am conscious of the conflicting roles I confronted while conducting the data collection, and at the same time working with the teachers in the PD in the department in the main study. The PD was conducted as action research (for a discussion on action research, see Section 3.3). The PD carried out in the study can be viewed as an action research because:

- It is a systematic investigation of assessment practice within the English department.
- It involves the collaboration of teachers in the department.
- It aimed to enhance teachers' understandings of performance-based assessment.
- It focused on introducing changes in assessment practices in the department.

In addition, the present study resembles an action research model because it is contextual, small-scale and localised, the main purpose is to bring about change and improvement in practices, it is a collaborative investigation by a team of teachers and a researcher, and the changes in practice originated from the data provided by the teachers.

Furthermore, the present study shares some features of the collaborative teacher development because PD is its prime purpose, teachers who participated in the PD had control over the process, and PD was to be built into the processes as its core component. The significant fundamental challenge I encountered was not on the power imbalance, but the conflicting roles while providing PD workshop and collecting data for the present research project (cf. Johnston, 2009, Section 3.3). There were three roles I was performing: the researcher, whose main purpose was to collect the data; the PD trainer who introduced changes in assessment to the department by enhancing teachers' understanding of performance-based assessment; and a member of staff who wanted to see the improvement in assessment practices in the department.

In retrospect, the role conflicts took place in the interviews with the PD participants. As the PD trainer in an action research, when the aim is to collaboratively improve assessment practices, when I found out that the beliefs or practices reported by the participants were not in line with assessment principles, I should have discussed them in the interviews because, as a PD trainer as well as a colleague, I would want to lead the participants in the right direction. However, as a

researcher, I would not want to interrupt the flow of the conversation in the interviews, or, cause interviewees to feel uncomfortable by telling them that they were wrong. In the present study, I decided to intervene when I believed that what the interviewees reported could have the effects on the wider community. For example, when Tanya reported about her involvement in preparing a standardisation meeting, I intervened by giving her some advice. I decided to step in because I believed that the standardisation meeting would affect more than 50 teachers who would be participating in this standardisation meeting. When I transcribed and analysed this interview, the question I had for myself was whether my intervention was appropriate.

I do not have a definite answer to this question, and a definite answer may not be possible. As a member of staff at the University, I believe that my action was justified as it was for the improvement of the assessment problems in the department, a goal which was likewise one of the main goals of the present study. Nevertheless, I did not intervene when the participants reported their personal beliefs which were not in line with assessment principles. I did not interfere when I believed what they reported did not have immediate effects on other teachers. For instance, when Songsri used the term 'washback' incorrectly, I did not tell her that she did not understand the concept of washback and used the term incorrectly. As a PD trainer, I should have provided some input for Songsri about washback; on the other hand, as a researcher conducting an interview, I was aware that Songsri might get intimidated by the remarks and feel uncomfortable sharing her thoughts and experiences with me in the future. Thus, I decided not to intervene in this circumstance. In any future action research or professional development, I would be more conscious of my multiple roles, and, develop strategies for performing all of them well at appropriate moment. For example, I could have arranged to chat with Songrsi later about washback in the context of the Department, and introduce a simple definition into the conversation; or

I could have prepared some handouts for interested teachers with key terms to help them learn to speak properly about assessment and its role in our teaching.

Although there might be some conflicts of roles, when an action research is being conducted by a researcher, who also a PD trainer and a member of staff, I may have made mistakes in balancing my roles. Nevertheless, I believe that the collaboration between a researcher (or researchers) and teachers in conducting action research can contribute to positive changes in individual teachers, as well as to a programme. I believe that the findings from the present study reveal this to be the case.

8.4 Limitations

The first limitation of my research pertains to the research design. First of all, because the present study only investigated the impact of the PD on teachers, I did not include any learners. In retrospect, I believe that it is important for future studies to include learners in the research design because the data from a students' point of view could reveal other aspects of impact. Mixing quantitative research methods, such as survey questionnaires for students, could also be implemented. Furthermore, concerning the research methods employed, the participants pointed out how think-aloud had an impact on how they rated students' performances. They stated that because of this research method, they had to rate the performances according to the criteria, which they might have done differently without the think-aloud.

The second limitation is the presentation of the data in the thesis. Although I collected a wealth of data to support the claims made in the thesis, space and time limitations inherent in a PhD project did not allow me to present all the data, and go into as much depth in presenting them, as might have been possible and preferable. For example, I did not include the data from the field observations from my involvement with assessment activities in the Department (such as the standardisation

meetings and the revision of the rating criteria projects) which suggested that the present study had directly affected the assessment practice in the department. However, this rich and untapped data will, I believe, be a useful resource for my future publications. It will also be of value in carrying out a longitudinal study of the impact of the PD because, as pointed out by many researchers, the process of change is long, complex, affected by many factors, and the rate of change is different for each individual. Only with further study and observation within the Department will it become clear what impact, if any, the PD and the present study will have had.

The third limitation concerns the nature of the actual PD itself. For the PD workshop, the main limitations were time constraints and the availability of the participants. It is known that language teachers have a lot to do on a day to day basis, having them involved in a PhD project, which requires a lot of their participation as in this study, was demanding. Some of my early intentions could not be carried out since the participants did not agree to participate because of time constraints. For instance, think-aloud protocol was intended to be conducted in the follow-up study to investigate the changes in the participants' rating practice. However, the participants expressed that they were not willing to do it in that phase of the research.

References

Alderson, J. C. (1991). Bands and scores. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s: The communicative legacy* (pp. 71-86). London: Macmillan.

Alderson, J. C. (2004). Forward. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. iv-xii). Mahwah, NJ: Lawrence Erlbaum Associates.

Alderson, J. C., & Banerjee, J. (2001). Language testing and assessment (Part 1). *Language Teaching, 34*(4), 213-236.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*(2), 115-129.

Alderson, J. C., & Wall, D. (Eds.). (1996). *Language Testing, 13*(3).

American Federation of Teachers National Council on Measurement in Education and the National Education Association. (1990). Standards for teacher competence in educational assessment of students. Retrieved March,17, 2008, from http://www.unl.edu/buros/bimm/html/article3.html

Andrews, S. (2004). Washback and curriculum innovation. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 37-50). Mahwah, NJ: Lawrence Erlbaum Associates.

Arter, J., & McTighe, J. (2001). *Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance*. Thousand Oaks, CA: Corwin Press.

Bacha, N. (2001). Writing evaluation: What can analytic versus holistic essay scoring tell us? *System, 29*(3), 371-383.

Bachman, L. F. & Palmer, A. S. (1996). Language testing in practice. Oxford: Oxford University Press.

Bailey, K. M. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing, 13*(3), 257-279.

Bailey, K. M. (1999). *Washback in language testing (TOEFL Monograph Series RM-99-4)*. Princeton, NJ: Educational Testing Service.

Baker, W. (2008). A critical examination of ELT in Thailand: The role of cultural awareness. *RELC 39*(1), 131-146.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing, 12*(1), 86-107.

Barrett, S. (2001). The impact of training on rater variability. *International Education Journal, 2*(1), 49-58.

Borg, S. (2003). Teacher cognition in language teaching: a review of research on what language teachers think, know, believe and do. *Language teaching, 36*(2), 81-109.

Borg, S. (2006). *Teacher cognition and language education*. London: Continuum.

Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. *Educational researcher, 33*(8), 3-15.

Brewster, S., Davies, P., & Rogers, M. (2001). *Skyline 3*. Oxford: Macmillan.

Brindley, G. (2001). Language assessment and professional development. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. McNamara & K. O'Loughlin (Eds.), *Experimenting with uncertainty: Essays in honour of Alan Davies* (Vol. 11, pp. 137-143). Cambridge: Cambridge University Press.

Brindley, G. (2008). Educational reform and language testing. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed., Vol. 7, pp. 365-378). New York, NY: Springer Science+Business Media.

Brown, A. (1995). The effect of rater variables in the development of an occupation - specific language performance test. *Language Testing, 12*(1), 1-15.

Brown, A. (2000). An investigation of the rating process in the IELTS oral interview. In R. Tulloh (Ed.), *IELTS Research reports 2000* (Vol. 3, pp. 49-84). Canberra: IETLS Australia Plt Limited.

Brown, J. D. (1998). *New ways of classroom assessment*. Bloomington: TESOL.

Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly, 32*(4), 653-675.

Burns, A. (1992). Teacher beliefs and their influence on classroom practice. *Prospect, 7*(3), 56-66.

Burns, A. (1996). Starting all over again: From teaching adults to teaching beginners. In D. Freeman & J. C. Richards (Eds.), *Teacher learning in language teaching* (pp. 154-177). Cambridge: Cambridge University Press.

Burns, A. (1999). *Collaborative action research for English language teachers.* Cambridge: Cambridge University Press.

Burns, A. (2005a). Action research: an evolving paradigm? *Language Teaching, 38*(2), 57-74.

Burns, A. (2005b). Action research. In E. Hinkel (Ed.), Handbook of research in second language teaching and learning (pp. 241-256). Mahwah, N.J.: Lawrence Erlbaum.

Burns, A. (2009). Action research in second language teacher education. In A. Burns & J. C. Richards (Eds.), *The Cambridge guide to second language teacher education* (pp. 289-297). Cambridge: Cambridge University Press.

Burrows, C. (2004). Washback in classroom-based assessment: A study of the washback effect in the Australian adult migrant English programme. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 113-128). Mahwah, NJ: Lawrence Erlbaum Associates.

Charmaz, K. (2002). Qualitative interviewing and grounded theory analysis. In J. F. Gubrium & J. A.Holstein (Eds.), *Handbook of interview research: Context & method* (pp. 675-694). Thousand Oaks, CA: Sage.

Charmaz, K. (2004). Grounded theory. In S. Hesse-Biber & P. Leavy (Eds.), *Approaches to qualitative research: a reader on theory and practice* (pp. 496-521). New York, NY: Oxford University Press.

Charmaz, K. (2006). *Constructing grounded theory*. London: Sage.

Cheng, L. (1997). How does washback influence teaching? Implications for Hong Kong. *Language and Education, 11*(1), 38-54.

Cheng, L. (1998). Impact of a public English examination change on students' perceptions and attitudes toward their English learning. *Studies in Education Evaluation, 24*(3), 279-301.

Cheng, L. (1999). Changing assessment: Washback on teacher perceptions and actions. *Teaching and teacher education, 15*(3), 253-271.

Cheng, L. (2005). *Changing language teaching through language testing: A washback study*. Cambridge: Cambridge University Press.

Cheng, L. (2008). Washback, impact and consequences. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed., Vol. 7, pp. 349-364). New York, NY: Springer Science+Business Media.

Cheng, L., & Curtis, A. (2004). Washback or backwash: A review of the impact of testing on teaching and learning. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in language testing: research contexts and methods* (pp. 3-17). Mahwah, NJ: Lawrence Erlbaum Associates.

Cheng, L., Todd, R., & Hu, H. (2004). ESL/EFL instructors' classroom assessment practices: purposes methods, and procedures. *Language Testing, 21*(3), 360-389.

Cizek, G., Fitzgerald, R., & Rachor, R. (1995). Teacher's assessment practices: Preparation, isolation and the kitchen sink. *Educational Assessment, 3*(2), 159-179.

Clarke, D., & Hollingsworth, H. (2002). Elaborating a model of teacher professional growth. *Teaching and teacher education, 18*(1), 957-967.

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education* (6th ed.). London: RoutedgeFalmer.

Commisson on Higher Education. (n.d.). Higher Education in Thailand. Retrieved October, 28, 2008, from http://www.kpi.mua.go.th/intcoop/main2/files/file/publications/book_higher_education/book_higher.pdf

Corbin, J., & Strauss, A. (2008). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (3rd ed.). Thousand Oaks, CA: Sage.

Crandall, J. (2000). Language teacher education. *Annual Review of Applied Linguistics, 20*(1), 34-55.

Creswell, J. W. (2009). *Research design: Qualitative, qualitative, and mixed methods approaches*. Thousand Oaks, CA: Sage.

Creswell, J. W. (2007). *Qualitative inquiry and research design: Choosing among five approaches* (2nd ed.). Thousand Oaks: CA: Sage.

Cumming, A., Kantor, R., & Powers, D. E. (2001). *Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: an investigation into raters' decision making and development of a preliminary analytic framework (Monograph Series MS-22)*. New Jersey: Educational Testing Service.

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision Making while Rating ESL/EFL Writing Tasks: A Descriptive Framework. *The Modern Language Journal, 8*(1), 67-96.

Davies, A. (2008). Textbook trends in teaching language testing. *Language Testing, 25*(3), 327-347.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing* (Vol. 7). Cambridge: Cambridge University Press.

Davison, C. (2004). The contradictory culture of teacher-based assessment: ESL teacher assessment practices in Australian and Hong Kong secondary schools. *Language Testing, 21*(3), 305-334.

Davison, C. (2007). Views from the chalkface: English language school-based assessment in Hong Kong. *Language Assessment Quarterly, 4*(1), 37-68.

Day, C., Sammons, P., & Gu, Q. (2008). Combining qualitative and quantitative methodologies in research on teacher's lives, work, and effectiveness: From integration to synergy. *Educational Researcher, 37*(6), 330-342.

Delamont, S. (2004). Ethnography and participant observation. In C. Seale, G. Gobo, J. F. Gubrium & D. Silverman (Eds.), *Qualitative research practice* (Concise Paperback ed., pp. 205-217). London: Sage.

Dey, I. (2004). Grounded theory. In C. Seale, G. Gobo, J. F. Gubrium & D. Silverman (Eds.), *Qualitative research practice* (Concise Paperback ed., pp. 80-93). London: Sage.

Dörnyei, Z. (2005). *The psychology of the language learner: Individual differences in second language acquisition.* Mahwah: Lawrence Erlbaum.

Dörnyei, Z. (2007). *Research methods in applied linguistics.* Oxford: Oxford University Press.

Douglas, D. (2001). Language for specific purposes assessment criteria: Where do they come from? . *Language Testing, 18*(2), 171-185.

Douglas, D., & Myers, R. (2000). Assessing the communication skills of veterinary students: Whose criteria? In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (pp. 60-81). Cambridge: Cambridge University Press.

Duff, P. A. (2007). *Case study research in applied linguistic*. New York: Lawrence Erlbaum Associates.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*(2), 188-185.

Edge, J. (2002). *Continuing cooperative development: A discourse framework for individuals as colleagues*. Ann Arbor, MI: University of Michigan.

Elder, C., Barkhuizen, G., Knoch, U., & Randow, J. v. (2007). Evaluating rater responses to an online training programme for L2 writing assessment. *Language Testing, 24*(1), 37-64.

English Department. (2006). *Teacher's guide: English 103 & 104*. Unpublished manuscript, Chiang Mai University, Chiang Mai, Thailand.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis* (Revised ed.). Massachusetts: Massachusetts Institute of Technology.

Fazio, R. H., & Olson, M. A. (2003). Attitude: Foundations, functions, and consequences. In M. A. Hogg & J. Coopper (Eds.), *The sage handbook of social psychology* (pp. 140-160). London: Sage.

Flyvbjerg, B. (2004). Five misunderstandings about case-study research. In C. Seale, G. Gobo, J. F. Gubrium & D. Silverman (Eds.), *Qualitative research practice* (Concise Paperback ed., pp. 390-404). London: Sage.

Foley, J. A. (2005). English in ... Thailand. *RELC, 36*(2), 223-234.

Fox, J. (2008). Alternative assessment. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education: language testing and assessment* (2nd ed., Vol. 7, pp. 97-109). New York: Springer Science+Business Media.

Freeman, D. (1989). Teacher training, development, and decision making: a model of teaching and related strategies for language teacher education. *TESOL Quarterly, 23*(1), 27-45.

Freeman, D. (1993). Renaming experience/ reconstructing practice: developing new understanding of teaching. *Teaching and teacher education,* 9(5/6), 485-497.

Fullan, M. (2007). *The new meaning of educational change* (4th ed.). New York: Teachers College Press, Columbia University.

Genesee, F., & Upshur, J. A. (1996). *Classroom-based Evaluation in Second Language Education*. Cambridge: Cambridge University Press.

Gillham, B. (2000). *The research interview*. London: Continuum.

Goulden, N. R. (1994). Relationship of analytic and holistic methods to raters' scores for speeches. *The Journal of Research and Development in Education, 27*(1), 73-82.

Guskey, T. R. (2000). *Evaluating professional development*. Thousand Oaks, CA: Corwin Press.

Guskey, T. R. (2002). Professional development and teacher change. *Teacher and Teaching: Theory and Practice, 8*(3/4), 381-391.

Hamp-Lyons, L. (Ed.). (1991a). *Assessing second language writing in academic context*. Norwood, NJ: Ablex Publishing Corporation.

Hamp-Lyons, L. (1991b). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic context* (pp. 241-278). Norwood, NJ: Ablex Publishing Corporation.

Hamp-Lyons, L. (1997a). Washback, impact and validity: Ethical concerns. *Language Testing, 14*(3), 295-303.

Hamp-Lyons, L. (1997b). Ethics in language testing. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education* (1st ed., Vol. 7, pp. 321-333). Dordrecht, the Netherlands: Kluwer Academic Publishers.

Hamp-Lyons, L. (1995). Rating nonnative writing: The trouble with holistic scoring. *TESOL Quarterly, 29*(4), 759-762.

Hamp-Lyons, L. (2001). Ethics, fairness(es), and developments in language testing. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. McNamara & K. O'Loughlin (Eds.), *Experimenting with uncertainty: essays in honour of Alan Davies* (Vol. 11, pp. 222-227). Cambridge: Cambridge University Press.

Hamp-Lyons, L. (2002). Editorial: The scope of writing assessment. *Assessing Writing, 8*(1), 5-16.

Hamp-Lyons, L. (2003). Writing teachers as assessors In B. Kroll (Ed.), *Exploring they dynamics of second language writing* (pp. 162-189). Cambridge: Cambridge University Press.

Hamp-Lyons, L. (2007a). Editorial: Worrying about rating. *Assessing Writing, 12*(1), 1-9.

Hamp-Lyons, L. (2007b). The impact of testing practices on teaching: Ideologies and alternative. In J. Cummins & C. Davison (Eds.), *International handbook of English language teaching* (Vol. Part I, pp. 487-504). New York: Springer.

Hamp-Lyons, L., & Kroll, B. (1997). *TOEFL 2000 - writing: composition, community, and assessment*. Princeton, NJ: Educational Testing Service.

Hawkey, R. (2006). *Impact Theory and Practice: Studies of the IELTS test and Progetto Lingue 2000* Cambridge: Cambridge University Press.

Henrichsen, L. E. (1989). *Diffusion of innovations in English language teaching: The ELEC effort in Japan, 1956-1968*. Connecticut: Greenwood Press.

Hesse-Biber, S., & Leavy, P. (2006). *The practice of qualitative research*. Thousand Oaks, CA: Sage.

Hill, K. (1997). Who should be the judge? The use of non-native speakers as raters on a test of English as an international language. In A. Huhta, V. Kohonen, L. Kurki-Suonio & S. Luoma (Eds.), *Current developments and alternatives in language assessment: proceedings of LTRC 96* (pp. 275-290). Jyväskylä, Finland: University of Jyväskylä.

Hudson, T. (2005). Trends in assessment scales and criterion-referenced language assessment. *Annual Review of Applied Linguistics, 55*(1), 205-227.

Inbar-Lourie, O. (2008). Language assessment culture. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed., Vol. 7, pp. 285-299). New York: Springer Science+Business Media.

Iwashita, N., & Grove, E. (2003). A comparison of analytic and holistic scales in the context of a specific-purpose speaking test. *Prospect, 18*(3), 25-35.

Jacoby, S., & McNamara, T. (1999). Locating competence. *English for Specific Purposes, 18*(3), 213-241.

Johnson, K. E. (1994). The emerging beliefs and instructional practices of preservice English as a second language teacher. *Teaching & teacher education, (10)*4, 439-452.

Johnston, B. (2009). Collaborative teacher development. In A. Burns & J. C. Richards (Eds.), *The Cambridge guide to second language teacher education* (pp. 241-249). Cambridge: Cambridge University Press.

Kagan, D. M. (1992). Implications of research on teacher belief. *Educational Psychologist, 27*(1), 65-90.

Kubanyiova, M. (2009). Possible selves in language teacher development. In Z. Dörnyei, & Ushioda, E. (Eds.), *Motivation, language identity and the L2 self* (pp. 314-332), Bristol: Multilingual Matters.

Kubanyiova, M. (in press). *Understanding language teachers' conceptual change: An anatomy of failure*. Basingstoke: Palgrave Macmillan.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales *Language Testing, 26*(2), 275-304.

Knoch, U. (2007a). Do empirically developed rating scales function differently to conventional scales for academic writing? *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 5*(1), 1-36.

Knoch, U. (2007b). 'Little coherence, considerable strain for reader': A comparison between two rating scales for the assessment of coherence. *Assessing Writing, 12*(2), 108-128.

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing, 12*(1), 26-43.

Kunnan, A. J. (Ed.). (2000). *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida*. Cambridge: Cambridge University Press.

Leung, C. (2004). Developing formative teacher assessment: Knowledge, practice, and change. *Language Assessment Quarterly, 1*(1), 19-41.

Leung, C. (2005). Classroom teacher assessment of second language development: construct as practice. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 869-888). New Jersey: Lawrence Erlbaum.

Linacre, J. M. (1989-2008). FACTS: Rasch measurement computer program. Chicago, IL: MESA Press.

Lodico, M. G., Spaulding, D. T., & Voegtle, K. H. (2006). *Methods in educational research: From theory to practice* (2nd ed.). San Francisco, CA: Jossey Bass.

Lumley, T. (2000). *The process of the assessment of writing performance: The rater's perspective.* Unpublished PhD Thesis, University of Melbourne, Victoria.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*(3), 246–276.

Lumley, T., & McNamara, T. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*(1), 54-71.

Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.

Lynch, B. K. (2001a). The ethical potential of alternative language assessment. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. McNamara & K. O'Loughlin (Eds.), *Experimenting with uncertainty: Language Testing in honour of Alan Davies* (Vol. 11, pp. 228-239). Cambridge: Cambridge University Press.

Lynch, B. K. (2001b). Rethinking alternative assessment from a critical perspective. *Language Testing, 18*(4), 351-372.

Lynch, B. K. (2003). *Language assessment and programme evaluation*. Edinburgh: Edinburgh University Press.

Mackey, A., & Gass, S. M. (2005). *Second language research: Methodology and design*. New Jeseay: Lawrence Erlbaum Associates, Inc.

Macnaghten, P., & Myers, G. (2004). Focus groups. In C. Seale, G. Gobo, J. F. Gubrium & D. Silverman (Eds.), *Qualitative research practice* (Concise Paperback ed., pp. 65-79). London: Sage.

Malone, M. E. (2008). Training in language assessment. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed., Vol. 7, pp. 225-239). New York: Springer Science+Business Media.

Mann, S. (2005). The language teacher's development: State-of-the-art article. *Language Teaching, 38*, 103-118.

Markee, N. (1997). *Managing curricular innovation*. Cambridge: Cambridge University Press.

Markee, N. (2001). The diffusion of innovation in language teaching. In D. Hall & A. Hewings (Eds.), *Innovation in English language teaching: A reader* (pp. 118-125). London: Routledge.

Mavrommatis, Y. (1997). Understanding assessment in the classroom: phases of the assessment - the assessment episode. *Assessment in Education, 4*(3), 381-399.

McDonough, K., & Chaikitmongkol, W. (2007). Teachers' and learners' reactions to a task-based EFL course in Thailand. *TESOL Quarterly, 41*(1), 107-132.

McNamara, T. (1996). *Measuring second language performance.* London: Addison Wesley Longman Ltd.

McNamara, T. (1997). Performance testing. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education: language testing and assessment* (Vol. Language Testing and Assessment, pp. 131-139). Dordrecht: Kluwer Academic Publishers.

McNamara, T. (2000). *Language testing*. Oxford: Oxford University Press.

McNamara, T. (2001). Editorial: Rethinking alternative assessment. *Language Testing, 18*(2), 329-332.

McNamara, T. (2008). The socio-political and power dimensions of tests. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education: Language testing and assessment* (2nd ed., Vol. 7, pp. 401-414). New York, NY: Springer Science+Business Media.

McTighe, J., & Emberger, M. (2002). Teamwork on assessment creates powerful professional development. *Journal of Staff Development, 27*(1), 38-44.

Mertler, C. A. (2001). Designing scoring rubrics for your classroom [Electronic Version]. *Practical assessment, research & evaluation, 7*. Retrieved March, 17, 2008, from http://PAREonline.net/getvn.asp?v=7&n=25

Miller, S. I., & Fredericks, M. (1999). How does grounded theory explain? *Qualitative Health Research, 9*(4), 538-551.

Miller, W. L., & Crabtree, B. F. (2004). Depth interview. In S. Hesse-Biber & P. Leavy (Eds.), *Approaches to qualitative research: A reader on theory and practice* (pp. 185-202). New York, NY: Oxford University Press.

Morgan, D. L. (2004). Focus groups. In S. Hesse-Biber & P. Leavy (Eds.), *Approaches to qualitative research: A reader on theory and practice* (pp. 263-285). New York, NY: Oxford University Press.

Morrison, B., & Hamp-Lyons, L. (2007). Grounded theory research: increasing accountability and credibility through the use of the 'worked example'. *The International Journal of Interdisciplinary Social Sciences, 2*(3), 413-424.

Myford, C. M. & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part 1. *Journal of applied measurement, 4*(4), 386-422.

Myford, C. M. & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part 2. *Journal of applied measurement, 5*(2), 189-227.

Norris, J. M., Brown, J. D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments*. Honolulu: University of Hawai'i.

North, B., & Schneider, B. (1998). Scaling descriptors for language proficiency scales. *Language Testing, 15*(2), 217-263.

O'Loughlin, K. (2000). The impact of gender in the IELTS oral interview. In R. Tulloh (Ed.), *IELTS Research reports 2000* (Vol. 3, pp. 1-28). Canberra: IETLS Australia Plt Limited.

O'Loughlin, K. (2002). The impact of gender in oral proficiency testing. *Language Testing, 19*(2), 169-192.

Pajares, M. F. (1992). Teachers' beliefs and educational research: cleaning up a messy construct. *Review of educational research, 62*(3), 307-332.

Penny, J., Johnson, R. L., & Gordon, B. (2000). The effect of rating augmentation on inter-rater reliability: An empirical study of a holistic rubric. *Assessing Writing, 7*(2), 143-164.

Pillay, H. (2002). *Teacher development for quality learning: The Thailand education reform project.* (Consulting report prepared for Office of the National Education Commission and the Asian Development Bank), Queensland: Queensland University of Technology.

Pollitt, A., & Murray, N. L. (1996). What raters *really* pay attention to. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: selected papers from the 15th Language Testing Research Colloquium (LTRC), Cambridge and Arnhem* (pp. 74-91). Cambridge: Cambridge University Press.

Prapphal, K. (2008). Issues and trends in language testing and assessment in Thailand. *Language Testing, 25*(1), 127-143.

Rapley, T. (2004). Interviews. In C. Seale, G. Gobo, J. F. Gubrium & D. Silverman (Eds.), *Qualitative research practice* (Concise Paperback ed., pp. 15-33). London: Sage.

QSR International (2006). NVivo: Qualitative data analysis software, Version 7.

Rea-Dickins, P., & Scott, C. (Eds.). (2007a). *Assessment in Education, 14*(1).

Rea-Dickins, P., & Scott, C. (2007b). Editorial: Washback from language tests on teaching, learning and policy: Evidence from diverse settings. *Assessment in Education, 14*(1), 1-7.

Reed, D., & Cohen, A. (2001). Revising raters and ratings in oral language assessment. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. McNamara & K. O'Loughlin (Eds.), *Experimenting with uncertainty: Language Testing in honour of Alan Davies* (Vol. 11, pp. 82-96). Cambridge: Cambridge University Press.

Richards, J. C., & Farrell, T. S. C. (2005). *Professional development for language teachers*. Cambridge: Cambridge University Press.

Richards, J. C., Gallo, P. B., & Renandya, W. A. (2001). Exploring teachers' beliefs and the processes of change. *PAC journal, 1*(1), 41-64.

Richards, K. (2003). *Qualitative inquiry in TESOL*. Basingstoke: Palgrave Macmillan.

Richardson, V., & Anders, P. L. (1994). The study of teacher change. In J. Sikula (Ed.), *Handbook of research on teacher education* (pp. 102-119). New York: Macmillan.

Rogers, E. M. (2003). *Diffusion of innovations* (5th ed.). New York, NY: Free Press.

Ryen, A. (2004). Ethical issues. In C. Seale, G. Gobo, J. F. Gubrium & D. Silverman (Eds.), *Qualitative research practice* (Concise Paperback ed., pp. 218-235). London: Sage.

Sakaui, K., & Gaies, S. J. (2003). Beliefs and metaphors of a Japanese teacher of English. In P. Lalaja & A. M. F. Barcelos (Eds.), *Beliefs about SLA: new research approaches* (pp. 153-170). Dordrecht: Kluwer Academic Publishers.

Saville, N., & Hawkey, R. (2004). The IELTS impact study: Investigating washback on teaching materials. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in language testing: research contexts and methods* (pp. 73-94). Mahwah, NJ: Lawrence Erlbaum Associates.

Shaw, S. D., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing*. Cambridge: Cambridge University Press.

Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing, 18*(3), 303-325.

Shohamy, E. (2001). *The power of tests: a critical view of the uses of language tests*. Essex: Longman Pearson.

Shohamy, E. (2007). Language tests as policy tools. *Assessment in Education, 14*(1), 117-130.

Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test impact revisited: Washback effect over time. *Language Testing, 13*(3), 298-317.

Shohamy, E., Gordon, C. M., & Kraemer, R. A. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal, 76*(1), 27-33.

Song, B., & Caruso, I. (1996). Do English and ESL faculty differ in evaluating the essays of native English-speaking and ESL students? *Journal of Second Language Writing, 5*(2), 163-182.

Spolsky, B. (1995). *Measured Words: The Development of Objective Language Testing* Oxford: Oxford University Press.

Spolsky, B. (2008). Language assessment in historical and future perspective. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education: language testing and assessment* (2nd ed., Vol. 7, pp. 445-454). New York: Springer Science+Business Media.

Stiggins, R., & Conklin, N. (1992). *In teacher's hands: Investigating the practices of classroom assessment*. Albany, NY: State University of New York.

Stobart, G. (2003). Editorial: The impact of assessment: intended and unintended consequences. *Assessment in Education, 10*(2), 139-140.

Thomson, R., Plumridge, L., & Holland, J. (2003). Longitudinal qualitative research: A developing methodology. *International Journal of Social Research Methodology, 6*(3), 185-187.

Tsui, A. B. M. (2007). What shapes teachers' professional development? In J. Cummins & C. Davison (Eds.), *International handbook of English language teaching* (Vol. 2, pp. 1053-1066). New York: Springer.

Turner, C. (2001). The need for impact studies of L2 performance testing and rating: identifying areas of potential consequences at all levels of the testing cycle. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. McNamara & K. O'Loughlin (Eds.), *Experimenting with uncertainty: Language Testing in honour of Alan Davies* (Vol. 11, pp. 138-149). Cambridge: Cambridge University Press.

Turner, C. E. (2000). Listening to the voices of rating scale developers: Identifying salient features for second language performance assessment. *The Canadian Modern Language Review/La Revue cannadienne des langues vivantes, 56*(4), 555-584.

Turner, C. E. (2006). Professionalism and high-stake tests: Teachers' perspectives when dealing with educational change introduced through provincial exams. *TESL Canada Journal/Revue TESL du Canada 23*(2), 54-76.

Turner, C. E., & Upshur, J. A. (2002). Rating scales derived from student samples: Effects of the scale maker and the student sample on scale content and student scores. *TESOL Quarterly, 36*(1), 49-70.

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal, 49*(1), 3-12.

Vaughan, C. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic context* (pp. 111-125). Norwood, NJ: Ablex Publishing Corporation.

Wall, D. (1996). Introducing new tests into traditional systems: Insights from general education and from innovation theory. *Language Testing, 13*(3), 334-354.

Wall, D. (1997). Impact and washback in language testing. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education* (1st ed., Vol. 7, pp. 291-302). Dordrecht, the Netherlands: Kluwer Academic Publishers.

Wall, D. (2000). The impact of high-stakes testing on teaching and learning: Can this be predicted or controlled? *System, 28*(4), 499-509.

Wall, D. (2005). *The impact of high-stakes examinations on classroom teaching: a case study using insights from testing and innovation theory*. Cambridge: Cambridge University Press.

Wall, D., & Alderson, J. C. (1993). Examining washback: The Sri Lanka impact study. *Language Testing, 10*(1), 41-70.

Wall, D., & Horák, T. (2006). *The impact of changes in the TOEFL examination on teaching and learning in Central and Eastern Europe. Phase I: The baseline study (TOEFL Monograph No. MS-34)*. Princeton , NJ: Educational Testing Service.

Wall, D., & Horák, T. (2007). Using baseline studies in the investigation of test impact. *Assessment in Education, 14*(1), 99-116.

Wall, D., & Horák, T. (2008). *The impact of changes in the TOEFL examination on teaching and learning in Central and Eastern Europe. Phase 2: Coping with change (TOEFLiBT-05)*. Princeton , NJ: Educational Testing Service.

Watanabe, Y. (1996). Does grammar translation come from the entrance examination? Preliminary findings from classroom-based research. *Language Testing, 13*(3), 318-333.

Weigle, S. C. (1994). *Effects of training on raters of English as a second language composition: Quantitative and qualitative approaches.* Unpublished PhD Thesis, University of California, Los Angles.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263-287.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Wigglesworth, G. (2008). Task and performance based assessment. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed., Vol. 7, pp. 111-122). New York: Springer Science+Business Media.

Winitchaikul, K., Wiriyachitra, A., & Chaikitmongkol, W. (2002). Feasibility study for the fundamental English courses at Chiang Mai University. Unpublished menuscript, Chiang Mai University, Chiang Mai, Thailand.

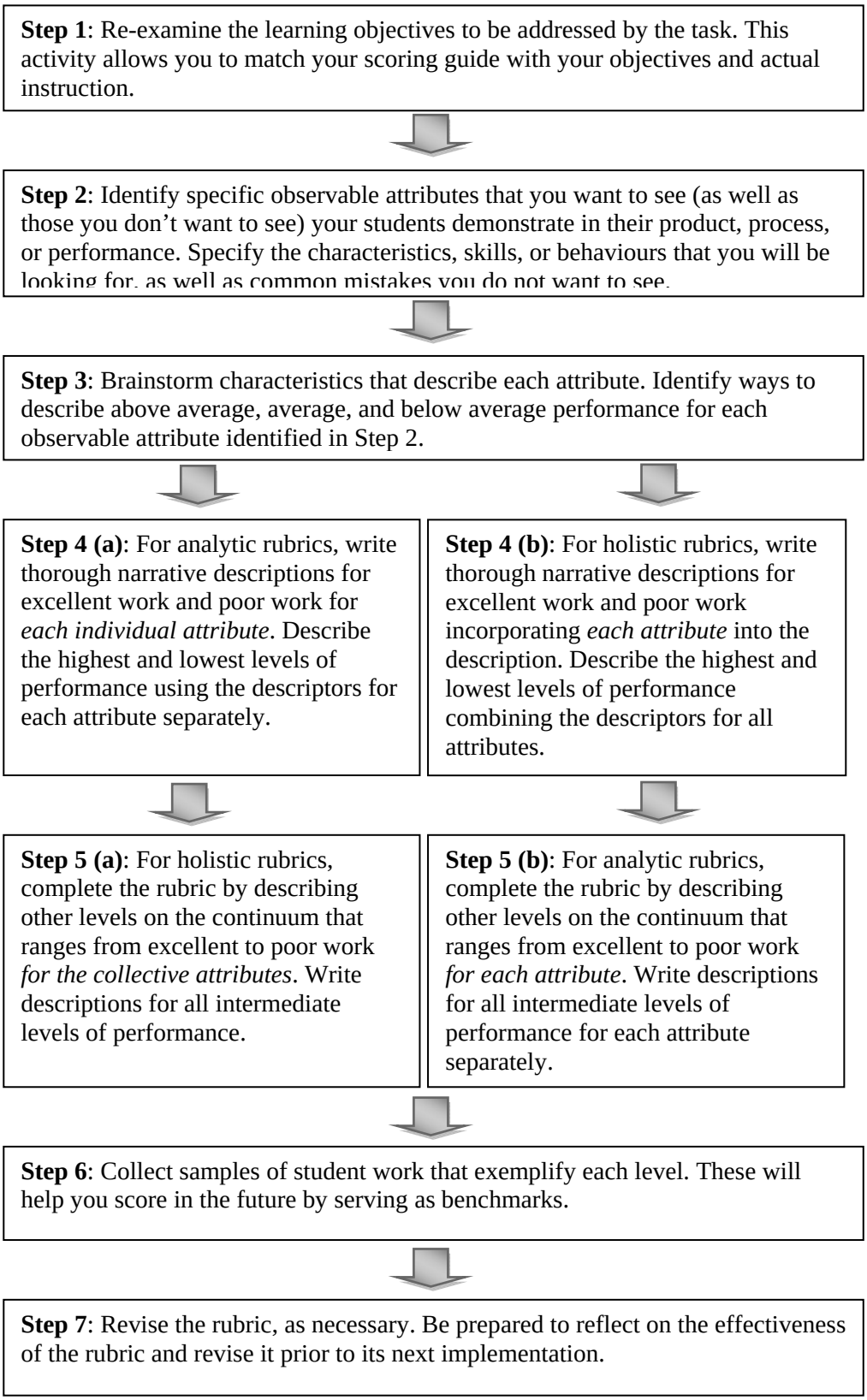
Wongsothorn, A., Hiranburana, K., & Chinnawongs, S. (2002). English language teaching in Thailand today. *Asia-Pacific Journal of Education, 22*(2), 107-116.

Woods, D. (1996). *Teacher cognition in language teaching: Beliefs, decision-making and classroom practice*. Cambridge: Cambridge University Press.

List of Appendices

APPENDIX A

Steps in designing rating scales



Adapted from Mertler (2001)

APPENDIX B

Standards for Teacher Competence in Educational Assessment of Students

1. Teachers should be skilled in choosing assessment methods appropriate for instructional decisions.

2. Teachers should be skilled in developing assessment methods appropriate for instructional decisions.

3. The teacher should be skilled in administering, scoring and interpreting the results of both externally-produced and teacher-produced assessment methods.

4. Teachers should be skilled in using assessment results when making decisions about individual students, planning teaching, developing curriculum, and school improvement.

5. Teachers should be skilled in developing valid pupil grading procedures which use pupil assessments.

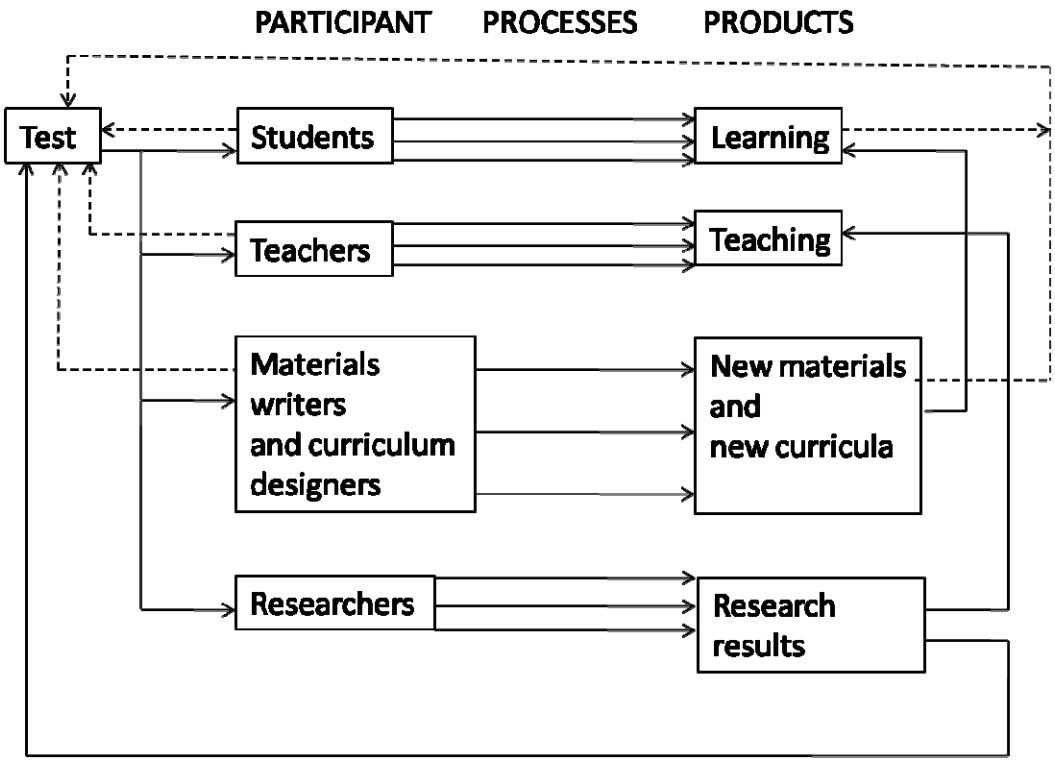
6. Teachers should be skilled in communicating assessment results to students, parents, other lay audiences, and other educators.

7. Teachers should be skilled in recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information.

(American Federation of Teachers National Council on Measurement in Education and the National Education Association, 1990)

APPENDIX C

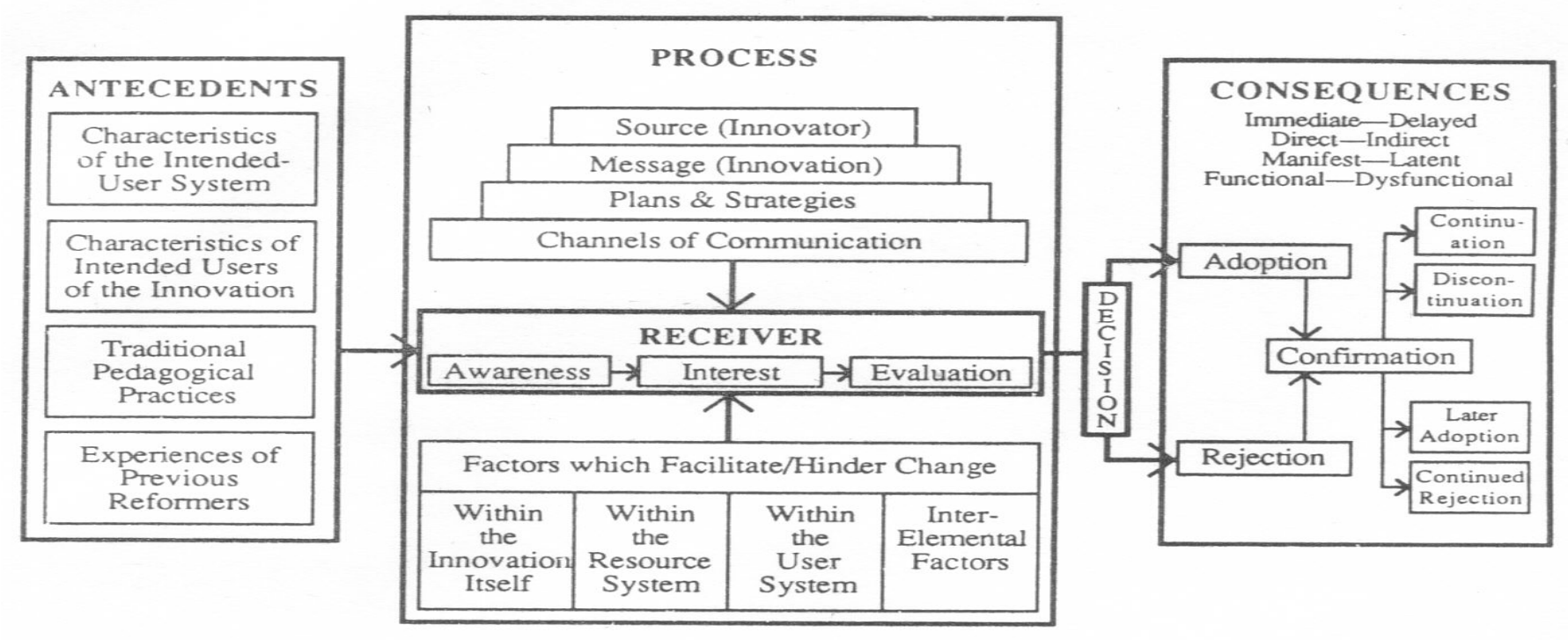
Basic model of washback



(Bailey, 1996, p. 264)

APPENDIX D

Henrichsen's hybrid model of diffusion/implementation process



(Henrichsen ,1989, p. 80)

APPENDIX E

Sample task and rating criteria

TASK 1: Travel Grants

Chiang Mai University is sponsoring overseas travel grants for CMU students to encourage educational tours and cultural understanding of countries around the world. The applicants in groups of three have to make an imaginary plan to visit only one foreign country (except England) for three days (not including departure and arrival days).

In order to get the grant, the applicants have to:
1. as a group, submit a three-day itinerary including interesting places they will visit as well as how to get there and travel around.
2. individually give an oral presentation describing one tourist attraction in the itinerary (two minutes for each person) including the following:
- name of the tourist attraction in the country he/she will visit and explain why he/she would like to go there,
- a description of that place including what & where it is, what to see & do, how to go, and time to go to that place, etc.

Grading Guideline for Written Itinerary

Very good (4 points)
- Nice layout
- Consisting of complete main elements:
  - Country, cities to visit
  - Number of days of visit
  - Departure and arrival dates and time
  - Description of each day: name(s) of city/cities and brief information of that day
- Places to visit match with group vacation profiles
- Appropriate use of emotive adjectives
- Mostly correct use of tenses
- Mostly correct use of spelling and grammar
- Having at least 6 – 7 sentences

Above average (3 or 3.5 points)
- Acceptable layout
- Including most of the main elements mentioned above
- Places to visit match with group vacation profiles
- Appropriate use of emotive adjectives
- Having a few mistakes in tenses
- Having a few mistakes in spelling and grammar
- Having at least 6 – 7 sentences

Average (2 or 2.5 points)
- Fair layout
- Missing some of the main elements mentioned above
- Places to visit do not clearly match with group vacation profiles
- Use of only a few appropriate emotive adjectives

- Having some mistakes in tenses
- Having some mistakes in spelling and grammar

Below average (1 or 1.5 points)
- Unorganized layout
- Missing many of the main elements mentioned above
- Places to visit do not match with group vacation profiles at all
- No use of appropriate emotive adjectives
- Having many mistakes in tenses
- Having many mistakes in spelling and grammar

Do not submit the itinerary (0)

Grading Guideline for Oral Presentation

	4 points	3 points	2 points	1 point	0
Content	- Describes the tourist attraction very clearly: 1. name of the attraction 2. reason(s) for visiting 3. description of that place: what it is where it is what to do how to go time to go	- Describes the tourist attraction clearly	- Includes some necessary information on the tourist attraction	- The tourist attraction is not clearly explained	
Language	- Very clear and correct pronunciation - Speaks fluently not reading the script all the time	- Clear and correct pronunciation - Look at the script sometimes	- Do not pronounce some words correctly - Often reads the script	- Has many problems with pronounce - Always reads the script	
Others	- Shows relevant picture(s) - Very good eye-contact	- Shows relevant picture(s) - Good eye-contact	- Shows relevant picture(s) - Eye-contact is not so good	- Does not show pictures - Eye-contact is not good enough	

APPENDIX F

Pilot study's interview schedules

- What do you know about assessment in general?
- Tell me about what assessments are used in FE 2.
- What was it like when you first experienced them? What did you think then?
- How would you describe how you viewed assessment before you employed FE 2 assessments? How has your view of assessment changed?
- How would you describe yourself as a teacher and assessor then?
- Tell me about your thoughts and feelings when you first exposed to FE 2 assessments?
- What happened next?
- Tell me about how you learned to handle FE 2 assessments?
- How have your thoughts and feelings about assessment changed since you first used the FE 2 assessments?
- What positive changes have occurred in your teaching since you first used the FE 2 assessments?
- What negative changes, if any, have occurred in your teaching since you first used the FE 2 assessments?
- Tell me how you go about this kind of assessment. What do you do?
- Could you describe a typical day for you when you teach? Now tell me about a typical day when you assess students.
- Where do you see yourself as a teacher and assessor in two years [five years, ten years, as appropriate]? Describe the person you hope to be then. How would you compare the person you hope to be and the person you see yourself as now?
- What helps you to manage English 104 assessments? What problems might you encounter? Tell me the sources of these problems.
- What do you think are the most important ways to assess students? How did you discover them? How has your experience before employing English 104 assessments affected how you handled assessment?
- Tell me about how your view and actions may have changed since you have been using FE 2 assessments?
- What advice would you give to someone who is new to FE 2 assessments?

APPENDIX G

Pilot study's consent Form

You are invited to be in a research study about your views toward FE 2 assessments. You will be asked interview questions about your views towards English 104 assessment and your classroom activities will be observed. Please read the form and ask questions you may have before agreeing to this study.

This study is being conducted by Bordin Chinda who is working on a PhD from the School of English Studies at the University of Nottingham, UK.

Background Information:

The purpose of this study is to conduct interviews and observations to be included in a thesis on teachers and language testing and assessment. The objectives of this study include:

- To understand the beliefs, knowledge and attitudes of Thai teachers at the Department towards the assessment being used and why they have those beliefs and attitudes
- To find out how teachers do assessment in their classroom and how their beliefs, knowledge and attitudes affect their practice in assessment
- To find out the areas in assessment which are needed to allow these teachers to improve their practice in assessment

Procedures:

If you agree to be in this study, you will be asked to do the following:

1. Agree to answer interview questions about your views towards English 104 assessment. There will be 2 interviews at the beginning and the end of the study.
2. Spend about 20 - 45 minutes with the researcher to answer these questions.
3. Agree to classroom observation.
4. Agree to be audio recorded.

Confidentiality:

The records of this study will be kept private. The information will be kept in a locked file for five years. While the information may be published, you will not be identified and your personal results will remain confidential. The researcher will be the only person who has access to the information gained during the study.

Contacts and Questions:

The researcher conducting this study is Bordin Chinda. You may ask any questions you have now. If you have questions later, you may contact him at: Bordin Chinda, School of English Studies, The University of Nottingham, NG7 2RD, United Kingdom or by email at: aexbc@nottingham.ac.uk. Student supervisor is: Professor Liz Hamp-Lyons at email: lizhl@hkucc.hku.hk.

You will be given a copy of this form to keep for your records.

Statements of consent:

I have read the above information. I have asked questions and received answers. I consent to participate in this study.

Print name: _____ Date: _____

Signature: _____

Signature of Researcher: _____ Date: _____

APPENDIX H

Instructions for think-aloud tasks

Think-aloud protocols ask people to say everything they think about while they perform a task, with the aim of documenting and better understanding what you pay attention to and consider important when you do a task. The purpose of the think-aloud protocols for this study is to find out in as much detail as possible what you are thinking about, deciding, and doing while you rate a sample of English 104 tasks. The most important thing to emphasize is, say everything you are thinking about, and make certain this is recorded clearly onto the tape recorder.

I am going to ask you to rate a set of 5 sample writing tasks. I would like you to rate them in the usual way. However, there will be one important difference with this batch: as I have previously mentioned, I am conducting a study of the processes used by teachers when they rate students' writing tasks, and I would now like you to talk and think aloud as you rate these scripts, while this audio recorder records what you say.

First, you should identify each script by the student ID as you start to read and rate it. Then, as you rate each task, you should vocalise your thoughts, and explain why you give the scores you give.

It is important that you keep talking all the time, registering your thoughts all the time. If you spend time reading the script or the rating scale, then you should do that aloud also, so that I can understand what you are doing at that time.

Notes

- Keep talking, conveying your thoughts continuously, while you assess the tasks, from the initial point when you first see each task until you have completed rating it, and indeed until you rate the whole set of them.
- Speak continuously. Report fully, even what might seem trivial. Do not assume that others know what you are doing or thinking.
- Try to avoid speech fillers (i.e., uh, um) as much as possible. Try to use words instead, so that we can understand what your thoughts have been.
- Talk and make your assessment as naturally and as honestly as you can, according to what you usually do when you assess students' tasks.

APPENDIX I

Main study's participant Information Sheet

Background Information:

This study is being conducted by Bordin Chinda who is working on a PhD in language testing and assessment from the School of English Studies at the University of Nottingham, UK.

The purposes of this study are to implement a professional development programme in assessment and to understand how it affects teachers who participate in the study.

The objectives of this study include:

1. To understand teachers' beliefs, attitudes and knowledge in language assessment
2. To understand the relationships between the above constructs and what teachers do
3. To implement a professional development programme
4. To understand how the professional development programme affects the teacher participants

Procedures:

If you agree to be in this study, you will be asked to do the following:

1. Answer the interview questions about your views towards language assessment. There will be 3 three interviews during the study.
2. Spend about 15 - 45 minutes with the researcher to answer these questions, which will be audio recorded.
3. Participate in a focus-group, which will be audio recorded.
4. Spend about 30 – 60 minutes with the researcher and the other participants in the focus-group discussing issues in language assessment.
5. Participate in the professional development programme, which will be video and audio recorded. The programme will include approximately 6 – 8 sessions.
6. Spend about 60 – 90 minutes with the researcher and the other participants and actively involve in the programme.
7. Take part in practical activity involving rating samples of students' performances, and validating and creating rating criteria.
8. Fill in a short evaluation/feedback form after each session of the programme.
9. Spend about 15 – 30 alone and/or with the researcher to do self-report of the three in-class assessments.

10. Invite the researcher to observe in lessons in order to enable him to have a better understanding of the classroom context to which the assessment applies.

Confidentiality:

The records of this study will be kept private. The information will be kept in a locked file for five years. While the information may be published, you will not be identified and your personal results will remain confidential. The researcher will be the only person who has access to the information gained during the study. While the researcher's supervisor may see the data, it will have been made anonymous before the supervisor sees it.

Participation:

Your participation in this research is completely voluntary and you may withdraw from the research project at any stage without prejudice or negative consequences. If you decline to take part in the research, non-participation will not affect your status now or in the future.

There is no potential risks or harms, to you or to your students, in participating this research project.

Contacts and Questions:

The researcher conducting this study is Bordin Chinda. You may ask any questions you have now. If you have questions later, you may contact him at School of English Studies, The University of Nottingham, NG7 2RD, United Kingdom or by email at: aexbc@nottingham.ac.uk.

The student supervisor is Professor Liz Hamp-Lyons at email: lizhl@hkucc.hku.hk.

APPENDIX J

Main study's participant Consent Form

Project title: Teachers' views of language assessment and an in-service professional development

Researcher's name: Mr. Bordin Chinda

Supervisor's name: Professor Liz Hamp-Lyons

- I have read the Participant Information Sheet and the nature and purpose of the research project has been explained to me. I understand and agree to take part.
- I understand the purpose of the research project and my involvement in it.
- I understand that I may withdraw from the research project at any stage and that this will not affect my status now or in the future.
- I understand that while information gained during the study may be published, I will not be identified and my personal results will remain confidential.
- I understand that I will be video recorded during the professional development programme.
- I understand that I will be audio recorded during the interviews.
- I understand that data will be kept private. The information will be kept in a locked file for five years. While the information may be published, you will not be identified and your personal results will remain confidential. The researcher will be the only person who has access to the information gained during the study.
- I understand that I may contact the researcher or supervisor if I require further information about the research.

Signed …… (research participant)

Print name …… **Date** ……

Contact details

Researcher: Mr. Bordin Chinda, School of English Studies, The University of Nottingham, NG7 2RD, United Kingdom or by email at: aexbc@nottingham.ac.uk

Supervisor: Professor Liz Hamp-Lyons at email: lizhl@hkucc.hku.hk.

APPENDIX K

Interview schedules

Main study:

Interview 1

1. How would you describe how you viewed language assessment?
2. What do you think are the most important ways to assess students?
3. What was it like when you first experienced assessments used in FE courses? What did you think then?
4. How has your view of assessment changed?
5. Tell me about how you learned to handle English 104 assessments?
6. Tell me how you go about this kind of assessment. What do you do?
7. How would you describe you as a teacher and assessor?
8. How have your thoughts and feelings about assessment changed since you first used the English 104 assessments?
9. What helps you to manage English 104 assessments? What problems might you encounter? Tell me the sources of these problems.
10. What advice would you give to someone who is new to English 104 assessments?

Interview 2

1. What made you decided to participate the programme?
2. What did you expect from participating in the programme?
3. So far, has the programme meet your expectations?
4. If yes, in what way?
5. If no, in what way and how to improve?
6. Have you come up with other expectations? What are they?
7. What do you like or dislike about the programme?
8. Did you learn anything new from the programme? What are they?
9. Have your started to view language assessment differently? How?
10. Have you considered doing assessment in your class differently? How?

Interview 3

Show the interviewee all materials used in the workshops (including the actual and the revised criteria). Then conclude what we have done in all workshops.

1. Has the programme met your expectations? What/How/What?
2. What do you think about the PD?
3. What do you think about the rating criteria (the new ones and the old ones)?
4. As an English teacher, what are your strengths and weaknesses in general & assessment?
5. And how do you view yourself in the past, present and future?
6. So far, what have been problems in your career? How did you deal with them? What about Assessment?
7. Now that we have finished the programme, do you have any plan for your professional development?

Follow-up study:

Interview 1

1. What do you think are the most important ways to assess students?
2. What was it like when you first experienced assessments used in FE courses? What did you think then?
3. How has your view of assessment changed?
4. Tell me about how you learned to handle English 104 assessments?
5. How would you describe yourself as a teacher and assessor?
6. How have your thoughts and feelings about assessment changed since you first used the English 104 assessments?
7. What problems might you encounter?
8. What did you like or dislike about the PD?
9. Did you learn anything from the programme? What are they?
10. Have your started to view language assessment differently? How/ What?
11. Have your started to do language assessment differently? How/ What?
12. What do you think about the PD?
13. What do you think about the rating criteria (the new ones and the old ones)?
14. Now that we have finished the programme, do you have any plan for your professional development?

Interview 2

1. What were your reactions when you first saw the new assessment criteria?
    a. Format
    b. Content
    c. Rating scales
2. Could you compare/contrast these criteria with the ones we revised in the workshops?
3. When rating, did you have any problem following the criteria?
4. How did you solve the problems?
5. Did you follow the criteria?
6. Did you do anything different from what you did during the think-aloud sessions last semester?

APPENDIX L

Sample email exchange with participants

Dear (participant),

I hope all is well. I'm wondering if you could please have a look at the attached file. It's part of my thesis, which is my interpretation of the interviews with you. Could you please see if what I've written is an accurate, sufficient, etc, interpretation of our interviews. I've chosen from the interviews only the parts that I'd use in my thesis. Please feel free to let me know if there are any words/sentences/parts you'd like me to taken out OR if there's anything you want me to add. And if you want to change how your identity would appear in my thesis, please let me know. I really appreciate your help very much. Thank you very much.

Looking forward to hearing from you soon.

Best wishes,

Sample responses:

I've read the interview interpretation of my part in detail (every sentence!!) and found that everything is fine.

In general, the transcription is fine. Please see what I have marked, especially on the second page. You wrote: "In addition, she said that performance or language should be the main focus of the assessment not the content or the task fulfilment." For this phrase "the content or the task fulfilment", I'm not quite sure that it is exactly what I want to say. Maybe you could find another term for me. I actually want to say that language should be the main focus of the assessment not other skills such as presentation skills or something else which does not exactly involve in language skills.

APPENDIX M

PD's revised rating scale

	4	3	2	1
I. Components	**Includes all** the required elements: ➢ Country to visit ➢ Number of days of visit ➢ Departure and arrival dates and times ➢ Name(s) of city/cities of each day ➢ Picture(s) of the trip highlights of each day	**Misses one** required element	**Misses two** required elements	**Misses three or more** required elements
II. Content	**Includes all aspects**: ➢ Places to visit ➢ How to visit ➢ What to do ➢ Time to go AND **adequately** describes the daily activities to make a complete itinerary	**Includes all aspects** BUT **inadequately** describes the daily activities to make a complete itinerary **OR** **Misses one aspect** BUT **adequately** describes the daily activities to make a complete itinerary	**Misses one aspect** AND **inadequately** describes the daily activities to make a complete itinerary	**Misses more than one aspects**
III. Required language patterns	**Includes all** AND with **adequate** number of: ➢ Emotive adjectives ➢ Time markers ➢ Relative clauses	**Includes all** BUT with **inadequate** number of the required language patterns	**Misses one** AND with **inadequate** number of the required language patterns	**Misses more than one** of the required language patterns

		OR **Misses one** BUT with **adequate** number of the required language patterns		
IV. Accuracy of language	**Has almost no** inaccuracies in form and use of the following: ➢ Tenses ➢ Emotive adjectives ➢ Time markers ➢ Relative clauses ➢ Spelling	**Has some** inaccuracies in form and use of language	**Has a lot of** inaccuracies in form and use of language	**Has too many** inaccuracies in form and use of language

Assessment Record

Group # _____ Country to visit _____

Members: 1) _____ 2) _____ 3) _____

CRITERIA FOR THE AWARD OF MARKS	**TEACHER'S COMMENTS**
(Circle number for each domain) 1. Components 4 3 2 1 2. Content 4 3 2 1 3. Required language patterns 4 3 2 1 4. Accuracy of language 4 3 2 1 TOTAL: _____/20	

APPENDIX N

Excerpts from think-aloud protocols

Catbandit:

nice layout ... nice layout ... name of the country number of day ... departure arrival OK "Day one --- explore the sunny of Southern California" explore the sun of California by sunbathing ... explore the sun "Surfing is also an option for whom like to have fun" ตรงนี้ไม่เกี่ยว ... "After lunch --- San Pedro Bay" mm มี places to visit ... mm … how to visit, how to visit มีหรือเปล่า catch the bus OK ... what to do มี time to go, at nine ... after lunch mm ... we will ... emotive adjective มี mm ... magnificent ... wonderful, thrilling xx savory แปลว่าอะไร ... savory, breath taking uh

Papone:

ก่อนอื่นก็ต้องดู criteria ก่อนเนอะว่ามันเป็นไง OK Criteria ของการตรวจ itinerary … OK … OK อยู่ที่หน้า 22 แค่ 4 เต็ม "Nice layout – " OK อันนี้คือ 4 คะแนน ถ้า 3 2 1 ก็ลดหลั่นตามนั้น .5 ได้ OK แป๊ปหนึ่ง OK ประเทศแรกที่เลือกมาเป็น India

India itinerary 3 days OK "depart Bangkok arrive India depart India arrive Bangkok" OK มีรูปติดมาเรียบร้อย OK format ตามที่สอน departure and arrival date OK format OK ใช้ได้ ต่อไปทางด้าน content "Day one India Saturday" อุ้ยทำไมขึ้นมาเป็น India Day two ก็ India Day three ก็ India ต้องเป็น Day one แล้วต่อเป็นชื่อเมือง ใช้ไม่ได้แล้วนิ "After arriving – into a hotel" อุ้ยเหมือนกับที่สอนในหนังสือเป๊ะเลย "At 9 -" เพราะฉนั่น อันนี้ต้องแก้ Day one India สงสัยต้องเป็น Day one Delhi OK ต่อ "After lunch – where we can take a magnificent photographs" OK มีการใช้ emotive adjective ที่สอน "Our next – before to sleep" โอ้วไม่ได้ประโยคนี้ผิด ต้องแก้ grammar ผิดโครงสร้างก็ผิด The last activity for today is going to spar ต้องแก้เป็น is going to spar at the hotel to take a leisurely before to sleep เอ๊ะแปลว่าอะไร to relax OK to relax OK before sleeping OK OK ประโยคอื่นOK ละ

Wanwisa:

เดี๋ยว make sure กับ criteria ก่อนก็แล้วกัน xxxxx OK above average, average, ah อันละจะดูกลุ่มแรกละ Our China itinerary OK departure กับ arrival ถูกต้อง ทีแรกเป็นอะไรอะ"After arriving – a hotel" ตาม model เป๊ะเลย "at 10 am – Great Wall" ต้องมี the นะจ๊ะ "which is – over ridge of the wall" verb หายไปไหนจ๊ะ where we will walk "over the ridge of the wall – learn a little bit" about "Chinese" about นะจ๊ะไม่ใช่ for "from" from "our guide – a wonderful dinner of Chinese food" enjoy a wonderful Chinese dinner สิจ๊ะไม่ใช่ a wonderful dinner of Chinese food มี dinner แล้วไม่ต้องมี food อีก มี emotive ไหม มีการใช้ time marker ไหม มี OK มี which where หรือเปล่า มี OK มี 6-7 ประโยคไหม ดูซิ 1 1 2 3 อ้ามี 4-5 ประโยคเอง อันนี้ อันนี้เอาไปสัก 2.5 ดีไหมเนี้ยะ ขอนับอีกทีสิ 1 2 3 4 OK too short นะจ๊ะ

Tanya:

Section 086 กลุ่มนี้ไป New Zealand OK เริ่มจาก … เริ่มจาก … OK layout ดู layout content ก่อน OK มีบอกประเทศชัดเจน มี บอกว่าไปกี่วัน 3 วัน มีdeparture มี arrival มีครบ Depart Bangkok Arrive

Auckland Arrive Bangkok ah Arrive เขียนผิด แล้วก็ มีรูปภาพครบ มีวันที่ไปครบ 3 วัน layout กั้นหน้ากั้นหลังOKใช้ได้ มาดูที่วันเดินทาง Day one Auckland Day two Rotorua Day three Wellington ไม่ได้ highlight ไว้ชัดเจนเนอะแล้วก็ ที่นี้มาดูใน mmm เนื้อความเนอะ “After arriving – which” ah มีการใช้ which เขียนถูกด้วย “We will see panoramic” ah panoramic เขียนผิด “Well will enjoy – westhaven Mairna” ไม่ยอมเขียน ชื่อเฉพาะไม่ยอมเขียนตัวโต W ต้องเขียนตัวใหญ่ “comma home to – history” ah อันนี้ขยาย Marina เนอะ “home of thousand” อันนี้ copy มา ภาษา copyมาเป็น chunk chunk เลย ต่อไป “After lunch, - where we can do shopping” มีการใช้ relative pronoun ใช้ถูกด้วย ah “where we can shopping” แล้วก็ไม่มี ควรจะเป็น go shopping หรือ do shopping “with its histiric buildings that” ah มีการใช้ relative pronoun that ใช้ถูกด้วย “that have been” transformed เขียนผิดมันต้องเป็น transformed “into boutique - will showcase” run-on sentence ตรงนี้ OK “will show case New Zealand’s unique wildlife” มาดูรูป ดูเหมือนว่า highlight มันน่าจะอยู่ที่ Auckland Bridge มากกว่าแต่รูปที่เอามาใส่เป็นรูปของ water world ซึ่งก็ไม่ได้เป็น highlight ของวันนั้น

Songsri
ก็จะศึกษาถึง grading criteria ก่อนเพื่อทำการตกลงในหัวใจก่อนว่าเช่นอัน good ก็จะเป็น layout ดี ซึ่งส่วนใหญ่ก็ layout ดี แล้วก็มีcomponent ต่างๆเหล่านี้ ก็มี ประเทศ เมืองที่ไปเยี่ยม จำนวนวันที่ไป 1 2 3 “departure arrival dates and time” ข้างบนส่วนมาก็มีและก็ “description of each day” มีชื่อของเมือง “and brief information of that day” () “appropriate use of emotive adjective” แล้วก็มี “correct use of tenses” ก็ต้องเป็น future “emotive adjective” ที่สอนในห้อง “spelling and grammar mostly” มันส่วนมาก ได้ 4 “Having at least 6-7 sentences” ถ้า 3 หรือ 3.5 แสดงว่าสูงกว่า มาตรฐาน “Acceptable layout” “main elements mentioned above” … มี emotive adjective แต่ อ๋อ มี “few mistakes in tenses” xxx อันนี้ก็ 3 ถ้า 2 ก็ดู layout ดี ขาดหายบางอย่างขององค์ประกอบ แล้วก็ สถานที่ไปท่องเที่ยวไม่ไปกับ group profile “having some mistakes in tenses” ใช้ emotive adjective อย่างเหมาะสมน้อยมาก แล้วก็ mistakes in tenses มีการใช้ tense ผิด แล้วก็ “mistakes in spelling and grammar” ตัวสะกด และ grammar ผิดบ้าง ทีนี้ถ้าได้ 1 หรือ 1.5 ก็หมายถึงว่า layout ไม่ organized เลย ไม่ใส่ element ที่สำคัญ place ที่ไปไม่ match กับ group vocation profile เด็กไม่ใช้ no use of appropriate use of emotive adjective และก็ many mistakes … แล้วก็ ไม่ระวังการสะกดและ grammarเริ่มจาก Italy itinerary 3 วัน ออกจากเชียงใหม่ไปที่โรม

APPENDIX O

Samples of FE 2 Task 1 Course Materials

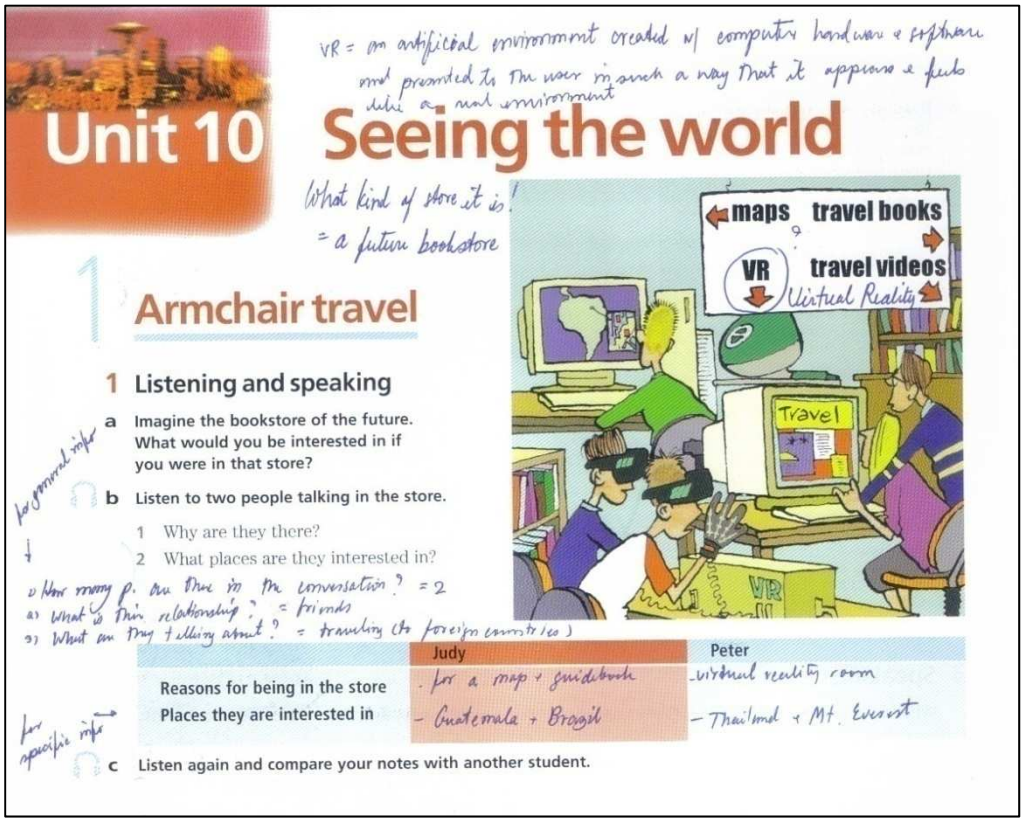
Task 1

Objectives

- Write a tour itinerary and give a presentation describing a tourist attraction
- Review the grammar items and vocabulary necessary for the task as well as reading and listening strategies learned in the previous course
- Introduces listening for general information, reading to identify the writer's purpose, and scanning strategies

Task 1 Class 1

- Introduce vocabulary and grammar related to the topic of travel
- Review some learning strategies learned in the previous course
    - Get students involved in the topic of travel
        - Vocabulary
        - Listening for general information
        - Listening for specific information
    - Introduce vacation profile
        - Reading to identify the writer's purpose
        - Making decisions from reading
    - Write their own vacation profiles
        - Vacation profile
    - Learn how to use unreal conditional sentences to express opinions on a country he/she would like to visit
        - Unreal conditional sentences



Skyline 3 (Brewster et al., 2001, p. 86).

APPENDIX P

Tentative timetables for the PD workshop and data collection activities

The workshops

Workshop 1	16 November	13.30 – 14.30
Workshop 2	23 November	13.30 – 14.30
Workshop 3	30 November	13.30 – 14.30
Workshop 4	14 December	13.30 – 14.30
Workshop 5	21 December	13.30 – 14.30
Workshop 6	28 December	13.30 – 14.30
Workshop 7	18 January	13.30 – 14.30
Workshop 8	1 February	13.30 – 14.30
Workshop 9	9 February	13.30 – 14.30

Individual Interviews

5 – 9 November

10 – 14 December

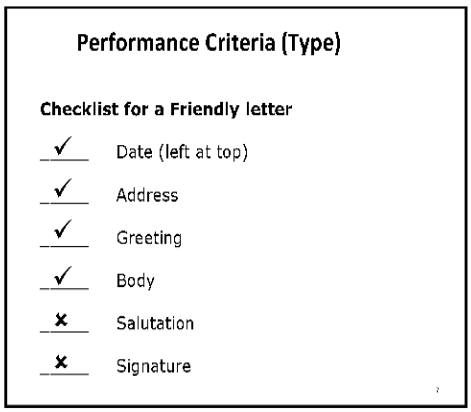
11 – 15 February

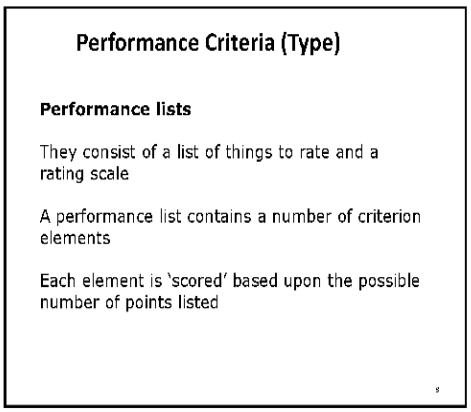
Think-aloud

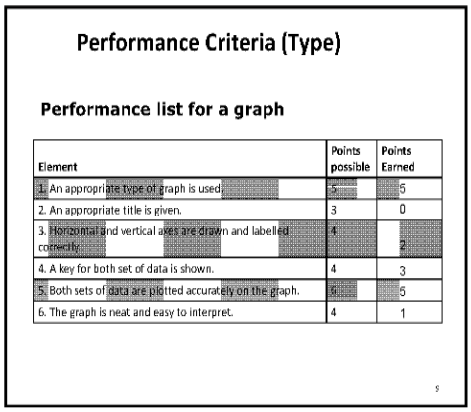
Task 1	26 – 30 November
Task 2	7 – 11 January
Task 3	4 – 8 February

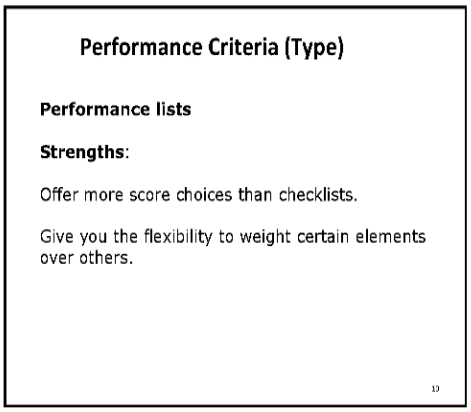
APPENDIX Q

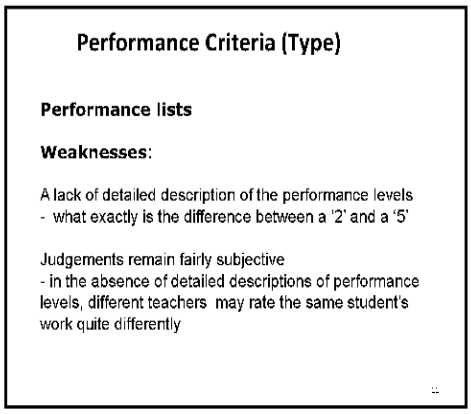
Sample slides of a Power Point Presentation on performance-based assessment

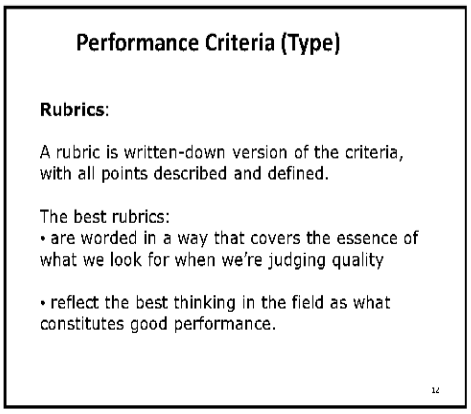








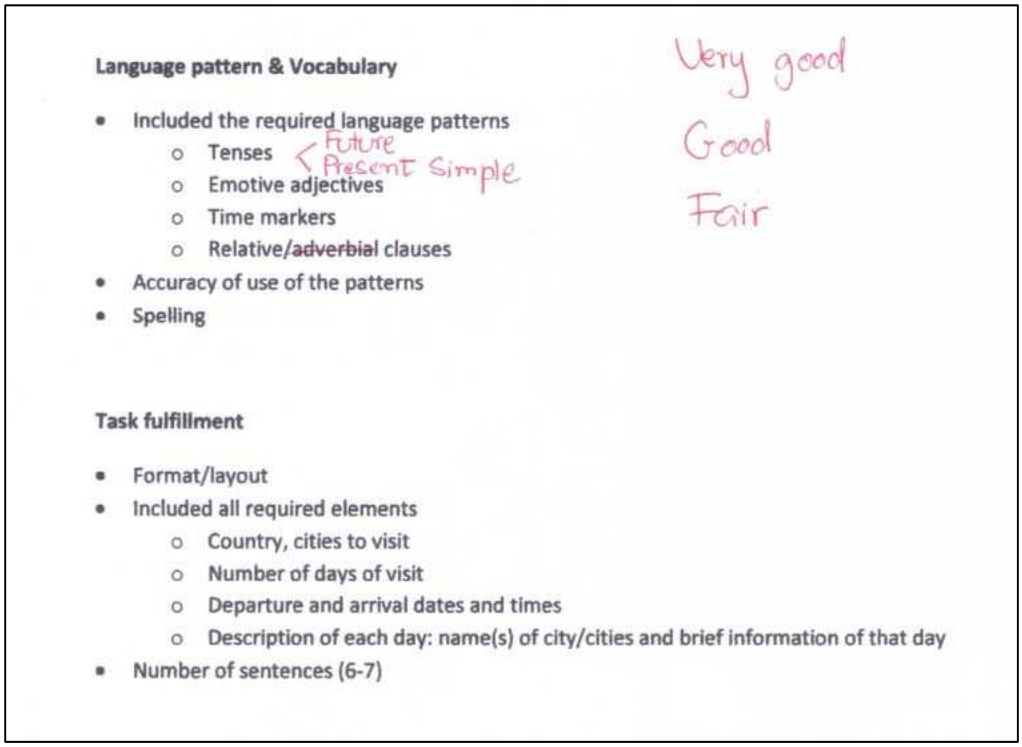




(from one of the papers presented during the data collection phase)

APPENDIX R

Samples of drafts of the rating criteria



Language pattern & Vocabulary

- Included the required language patterns
  - Tenses (Future, Present Simple)
  - Emotive adjectives
  - Time markers
  - Relative/~~adverbial~~ clauses
- Accuracy of use of the patterns
- Spelling

Very good
Good
Fair

Task fulfillment

- Format/layout
- Included all required elements
  - Country, cities to visit
  - Number of days of visit
  - Departure and arrival dates and times
  - Description of each day: name(s) of city/cities and brief information of that day
- Number of sentences (6-7)

Early draft of the rating criteria

	4	3	2	1
I. Format I. Components	**Includes all** the required elements ~~of the format:~~ 1. Country to visit 2. Number of days of visit 3. Departure and arrival dates and times 4. ~~Description of each day:~~ name(s) of city/cities ~~and brief information of that day.~~ 5. Good layout 6. Pictures of the trip highlights	**Misses one** required element ~~of the format~~	**Misses two** required elements ~~of the format~~	**Misses three or more** required elements ~~of the~~ format
II. Content	**Includes all aspects** and **adequately** describes the information needed: 1. Places to visit 2. How to visit 3. What to do 4. Time to go	**Includes all aspects** but **inadequately describes** the information needed OR **Misses one aspect** but **adequately** describes the information needed	**Misses one aspect** and **inadequately describes** the information needed	**Misses more than one aspects** of the information needed

Excerpt from the rating criteria after a few revisions

APPENDIX S

Sample of student's written performance (FE 2 Task 1)

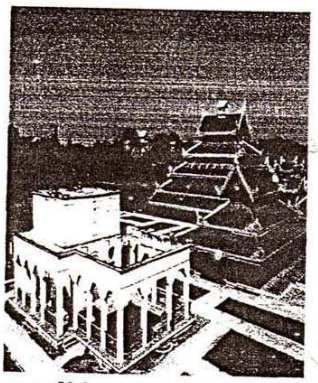
Our Myanmar Itinerary
3 days

Depart Chiang Mai: January 10th Time 5.00 a.m. Nok Air: NY101
Arrive Yangon: January 10th Time 7.00 a.m.
Depart Yangon: January 12th Time 8.00 p.m. Nok Air: NY111
Arrive Chiang Mai: January 12th Time 10.00 p.m.

Fair Langu
Good Task

Day 1 Yangon

After arriving in Myanmar in the early morning, we will go to Yangon which has a variety of interesting sectors such as ancient pagodas, beautiful parks, stunning images and museums. We will check into a Yangon hotels and have breakfast here. After breakfast we will tour around the city by bus for look along side views. In the afternoon we will arrive Shewdagon Pagoda where the capital city of Yangon and
The most scared pagoda as it enshrines the relics of the three earlier Buddha's and the gight hairs of Guatama Buddha. At this place we will have lunch and shopping along this day. In the evening we will back the Yangon hotel to take a bath and have dinner. About 10 p.m. we will go to bed and sleep well tonight.

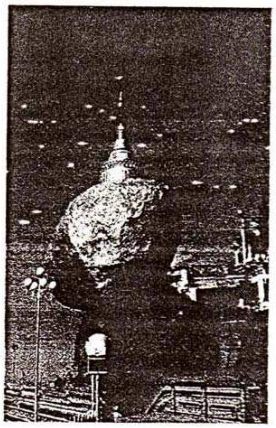


Day 2 Mandalay

After have breakfast in the early morning. We will start our scenic trip on a bus to Mandalay city. At 8a.m. we will going to the Royal Palace where the most fascinating Famous and amazing style of architecture, at here we will take the pick up and bring us to the foot of Mandalay and surrounding plains. In the evening we will back to Yangon hotel. We will take a bath about 2 hour after that we will go to the restaurant for dinner at 8 p.m. and go to bed at 10 p.m.

Day 3 Mon

We will fly to Heho Airport at Mon city, after breakfast when we get there, we will go to the famous Inlay Lake where southern Mon city in the eastern part of Myanmar. When arrive there we will go to Kyaikhtiyo Pagoda It is a 7.3 meters pagoda on top of a big "Golden Rock". The massive golden boulder is right on top of the Kyaikhto mountain.Legend say that the boulder maintains its precarius balance due to a precisely placed Buddha hair inside the pagoda. Once there was a King who received the Buddha's hair in the 11th century from a hermit. We will spend the whole afternoon in the beautiful lake. We can rest and take photo along this day. At 6 p.m. we will back to Yangon hotel by airplane we will take a bath and pack us bag about 2 hour. Later we will have a wonderful fare well dinner to celebrate our successful Myanmar vocation at this hotel after that we are going to Chiang Mai by airplane about 10 p.m.



APPENDIX T

Excerpt from Glossary of terms

Definitions of Related Terms

Analytic scoring: A method of subjective scoring often used in the assessment of speaking and writing skills, where a separate score is awarded for each of a number of features of a task, as opposed to one global score. In the assessment of writing the functional trisection of content, organisation and structure is commonly represented in the assessment categories. In speaking tests, commonly used categories are pronunciation or intelligibility, fluency, accuracy and appropriateness. Advantages claimed for the analytic method of scoring are that:

- raters are required to focus on each of the nominated aspects of performance individually, thus ensuring that they are all addressing the same features of the performance;
- it allows for more exact diagnostic reporting of literacy or oracy development, especially where skills my be developing at different rates (reflected in a marked profile);
- it leads to greater reliability as each candidate is awarded a number of scores.

A criticism commonly made of analytic scoring is that the focus on specified aspects of the performance may divert raters' attention from its overall effect. This problem may be at least partially overcome by requiring raters to give an overall impression score in addition to the analytic scores. A further problem with analytic scoring lies in the possibility of a halo effect distorting the score due to the number of judgement required. The main practical disadvantage of this method of scoring is that it is time consuming compared with holistic scoring.

Holistic/global scoring: A type of marking procedure which is common in communicative language testing whereby raters judge a stretch of discourse (spoken or written) impressionistically according to its overall properties rather than providing separate scores for particular features of the language produced (eg accuracy, lexical range). In the assessment of writing, a major advantage of holistic scoring over analytic scoring is that each piece of writing can be scored quickly, enabling each to be assessed by more than one rater for the same cost as one rater using several analytic criteria, thus leading, it is claimed, to greater reliability. A problem with holistic judgements, however, is that different raters may choose to focus on different aspects of the performance, leading potentially to poor reliability if only one rater is used. For the sake of reliability, therefore, test performance is normally judged by several raters and their judgements pooled. A further drawback of holistic scoring is that it does not allow detailed diagnostic information to be reported.

From Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. & McNamara, T. (1999) *Dictionary of Language Testing*, Cambridge: Cambridge University Press.

Appendix U

Excerpts from Samples assessment criteria

	I. Pronunciation & Delivery	**II. Communication Strategies**
6	**Can** project the voice appropriately for the context. **Can** pronounce all sounds/sound clusters and words clearly and accurately. **Can** speak fluently and naturally, with very little hesitation, and using intonation to enhance communication.	**Can** use **appropriate** body language to **display and encourage** interest. **Can** use **a full range** of turn-taking strategies to initiate and maintain **appropriate** interaction, and **can draw others** into extending the interaction (e.g. by summarising for others' benefit, or by redirecting a conversation); **can avoid the use of narrowly-formulaic expressions** when doing this.
5	**Can** project the voice appropriately for the context. **Can** pronounce all sounds/sound clusters clearly and almost all words accurately. **Can** speak fluently with only occasional hesitation, and using intonation to enhance communication, giving an overall sense of natural nonnative language.	**Can** use **appropriate** body language to **display and encourage** interest. **Can** use **a good range** of turn-taking strategies to **initiate and maintain** appropriate interaction (e.g. by encouraging contributions from others' in a group discussion, by asking for others' opinions, or by responding to questions); **can mostly avoid the use of narrowly-formulaic expressions** when doing this.

	I. Pronunciation & Delivery	**II. Communication Strategies**
6	**Can** project the voice appropriately for the context. **Can** pronounce all sounds/sound clusters and words clearly and accurately. **Can** speak fluently and naturally, with very little hesitation, and using intonation to enhance communication.	**Can** use appropriate body language to show focus on audience and to engage interest. **Can** judge timing in order to complete the presentation. **Can** confidently invite and respond to questions or comments when required for the task.
5	**Can** project the voice appropriately for the context. **Can** pronounce all sounds/sound clusters clearly and almost all words accurately. **Can** speak fluently with only occasional hesitation, and using intonation to enhance communication, giving an overall sense of natural nonnative language.	**Can** use appropriate body language to show focus on audience and to engage interest. **Can** judge timing sufficiently to cover all essential points of the topic. **Can** appropriately invite and respond to questions or comments when required for the task.

(Davison, 2007, pp. 61-66)